

Encyclopedia of
Complexity and Systems Science Series
Editor-in-Chief: Robert A. Meyers

SPRINGER
REFERENCE

Boris S. Kerner *Editor*

Complex Dynamics of Traffic Management

A Volume in the Encyclopedia of
Complexity and Systems Science,
Second Edition

 Springer

Encyclopedia of Complexity and Systems Science Series

Editor-in-Chief
Robert A. Meyers

The Encyclopedia of Complexity and Systems Science Series of topical volumes provides an authoritative source for understanding and applying the concepts of complexity theory, together with the tools and measures for analyzing complex systems in all fields of science and engineering. Many phenomena at all scales in science and engineering have the characteristics of complex systems and can be fully understood only through the transdisciplinary perspectives, theories, and tools of self-organization, synergetics, dynamical systems, turbulence, catastrophes, instabilities, nonlinearity, stochastic processes, chaos, neural networks, cellular automata, adaptive systems, genetic algorithms, and so on. Examples of near-term problems and major unknowns that can be approached through complexity and systems science include: the structure, history, and future of the universe; the biological basis of consciousness; the integration of genomics, proteomics, and bioinformatics as systems biology; human longevity limits; the limits of computing; sustainability of human societies and life on earth; predictability, dynamics, and extent of earthquakes, hurricanes, tsunamis, and other natural disasters; the dynamics of turbulent flows; lasers or fluids in physics; microprocessor design; macromolecular assembly in chemistry and biophysics; brain functions in cognitive neuroscience; climate change; ecosystem management; traffic management; and business cycles. All these seemingly diverse kinds of phenomena and structure formation have a number of important features and underlying structures in common. These deep structural similarities can be exploited to transfer analytical methods and understanding from one field to another. This unique work will extend the influence of complexity and system science to a much wider audience than has been possible to date.

More information about this series at <https://link.springer.com/bookseries/15581>

Boris S. Kerner
Editor

Complex Dynamics of Traffic Management

A Volume in the Encyclopedia of
Complexity and Systems Science,
Second Edition

With 408 Figures and 26 Tables

 Springer

Editor

Boris S. Kerner
Physics of Transport and Traffic
University Duisburg-Essen
Duisburg, Germany

ISBN 978-1-4939-8762-7 ISBN 978-1-4939-8763-4 (eBook)

ISBN 978-1-4939-8764-1 (print and electronic bundle)

<https://doi.org/10.1007/978-1-4939-8763-4>

Library of Congress Control Number: 2019935316

© Springer Science+Business Media, LLC, part of Springer Nature 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.

The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Series Preface

The Encyclopedia of Complexity and System Science Series is a multivolume authoritative source for understanding and applying the basic tenets of complexity and systems theory as well as the tools and measures for analyzing complex systems in science, engineering, and many areas of social, financial, and business interactions. It is written for an audience of advanced university undergraduate and graduate students, professors, and professionals in a wide range of fields who must manage complexity on scales ranging from the atomic and molecular to the societal and global.

Complex systems are systems that comprise many interacting parts with the ability to generate a new quality of collective behavior through self-organization, e.g., the spontaneous formation of temporal, spatial, or functional structures. They are therefore adaptive as they evolve and may contain self-driving feedback loops. Thus, complex systems are much more than a sum of their parts. Complex systems are often characterized as having extreme sensitivity to initial conditions as well as emergent behavior that are not readily predictable or even completely deterministic. The conclusion is that a reductionist (bottom-up) approach is often an incomplete description of a phenomenon. This recognition that the collective behavior of the whole system cannot be simply inferred from the understanding of the behavior of the individual components has led to many new concepts and sophisticated mathematical and modeling tools for application to many scientific, engineering, and societal issues that can be adequately described only in terms of complexity and complex systems.

Examples of Grand Scientific Challenges which can be approached through complexity and systems science include: the structure, history, and future of the universe; the biological basis of consciousness; the true complexity of the genetic makeup and molecular functioning of humans (genetics and epigenetics) and other life forms; human longevity limits; unification of the laws of physics; the dynamics and extent of climate change and the effects of climate change; extending the boundaries of and understanding the theoretical limits of computing; sustainability of life on the earth; workings of the interior of the earth; predictability, dynamics, and extent of earthquakes, tsunamis, and other natural disasters; dynamics of turbulent flows and the motion of granular materials; the structure of atoms as expressed in the Standard Model and the formulation of the Standard Model and gravity into a Unified Theory; the structure of water; control of global infectious diseases; and also evolution and quantification of (ultimately) human cooperative behavior in politics,

economics, business systems, and social interactions. In fact, most of these issues have identified nonlinearities and are beginning to be addressed with nonlinear techniques, e.g., human longevity limits, the Standard Model, climate change, earthquake prediction, workings of the earth's interior, natural disaster prediction, etc.

The individual complex systems mathematical and modeling tools and scientific and engineering applications that comprised the *Encyclopedia of Complexity and Systems Science* are being completely updated and the majority will be published as individual books edited by experts in each field who are eminent university faculty members.

The topics are as follows:

Agent Based Modeling and Simulation
 Applications of Physics and Mathematics to Social Science
 Cellular Automata, Mathematical Basis of
 Chaos and Complexity in Astrophysics
 Climate Modeling, Global Warming, and Weather Prediction
 Complex Networks and Graph Theory
 Complexity and Nonlinearity in Autonomous Robotics
 Complexity in Computational Chemistry
 Complexity in Earthquakes, Tsunamis, and Volcanoes, and Forecasting and
 Early Warning of Their Hazards
 Computational and Theoretical Nanoscience
 Control and Dynamical Systems
 Data Mining and Knowledge Discovery
 Ecological Complexity
 Ergodic Theory
 Finance and Econometrics
 Fractals and Multifractals
 Game Theory
 Granular Computing
 Intelligent Systems
 Nonlinear Ordinary Differential Equations and Dynamical Systems
 Nonlinear Partial Differential Equations
 Percolation
 Perturbation Theory
 Probability and Statistics in Complex Systems
 Quantum Information Science
 Social Network Analysis
 Soft Computing
 Solitons
 Statistical and Nonlinear Physics
 Synergetics
 System Dynamics
 Systems Biology

Each entry in each of the Series books was selected and peer reviews organized by one of our university- or industry-based book Editors with

advice and consultation provided by our eminent Board Members and the Editor-in-Chief.

This level of coordination assures that the reader can have a level of confidence in the relevance and accuracy of the information far exceeding that generally found on the World Wide Web. Accessibility is also a priority and for this reason each entry includes a glossary of important terms and a concise definition of the subject. In addition, we are pleased that the mathematical portions of our Encyclopedia have been selected by Math Reviews for indexing in MathSciNet. Also, ACM, the world's largest educational and scientific computing society, recognized our *Computational Complexity: Theory, Techniques, and Applications* book, which contains content taken exclusively from the *Encyclopedia of Complexity and Systems Science*, with an award as one of the notable Computer Science publications. Clearly, we have achieved prominence at a level beyond our expectations, but consistent with the high quality of the content!

Palm Desert, CA, USA
March 2019

Robert A. Meyers
Editor-in-Chief

Volume Preface

The complexity of vehicular traffic management is associated with the complex spatiotemporal dynamic behavior of vehicular traffic. The spatiotemporal traffic dynamic behavior is mostly associated with the occurrence of traffic breakdown, i.e., a transition from free flow to congested traffic in a traffic or transportation network. Congested traffic causes a considerable increase in travel time, fuel consumption, CO₂ emission, as well as other travel costs. For this reason, users of traffic and transportation networks expect that, through the application of traffic management, either traffic breakdown can be prevented in a network or, if the latter is not possible, the effect of traffic congestion on travel costs is reduced. Therefore, any traffic and transportation theory and model applied for traffic management should be consistent with empirical features of traffic breakdown at a network bottleneck.

Empirical traffic breakdown at a highway bottleneck is a transition from free traffic flow (F) to synchronized traffic flow (S) (called an F→S transition). Synchronized traffic flow is one of the phases of congested traffic. The most important empirical feature of traffic breakdown in a network is the nucleation nature of traffic breakdown found in real field traffic data. The term *nucleation nature* of traffic breakdown means that traffic breakdown occurs in a metastable free flow with respect to an F→S transition. The metastability of free flow is as follows. There can be many speed (density, flow rate) disturbances in free flow. Amplitudes of the disturbances can be very different. When a disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown (F→S transition) occurs. Such a disturbance resulting in breakdown is called the *nucleus* for the breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays; as a result, no traffic breakdown occurs.

To develop strategies and methods for traffic management, a diverse variety of traffic and transportation theories and models have been introduced and developed. These theories and models have had a great impact on the understanding of many empirical traffic phenomena observed in real traffic. Unfortunately, these generally accepted classical (standard) traffic and transportation theories and models have nevertheless failed by their applications in the real world. Even several decades of very intensive effort to improve and validate network optimization and control models based on the classical traffic and transportation theories have had no success. Indeed, no examples can be found where online implementations of the network optimization models based on

these classical traffic and transportation theories could reduce congestion in real traffic and transportation networks.

This failure of standard traffic and transportation theories can be explained as follows: The standard traffic and transportation theories are not consistent with the nucleation nature of traffic breakdown (F→S transition). The nucleation nature of empirical traffic breakdown was understood only during the last 20 years. In contrast, the generally accepted fundamentals and methodologies of traffic and transportation theory were introduced in the 1950s–1960s. Thus, the scientists whose ideas led to the classical (standard) traffic and transportation theories did not appreciate the empirical nucleation nature of traffic breakdown. Respectively, in the 1990s–2010s, the three-phase traffic theory and the breakdown minimization (BM) principle were introduced. The three-phase theory explains the empirical nucleation nature of traffic breakdown. The BM principle describes how to maximize the network throughput by preventing traffic breakdown in the network.

The standard traffic and transportation theories that failed when applied in the real world are currently the methodologies of teaching programs in most universities and the subject of publications in most transportation research journals and scientific conferences.

It could be assumed that the failure of standard methodologies of traffic and transportation science for reliable traffic management can be explained by a very long time interval between the development of the classical traffic theories made in the 1950s–1960s and the understanding of the empirical nucleation nature of traffic breakdown at road bottlenecks made at the end of the 1990s. During this long time interval, several generations of traffic researches developed a huge number of traffic flow models and strategies for traffic management based on the classical (standard) approaches. It turned out that the three-phase traffic theory is *incommensurable* with any classical (standard) traffic flow theory. The term *incomprehensibility* has been introduced by Kuhn in his famous book *The Structure of Scientific Revolutions* to explain a paradigm shift in a field of science. In particular, Kuhn's historical analysis shows that there is some common behavior of a research community during a paradigm shift. From Kuhn's analysis, it is understandable why it is very difficult for most members of the traffic and transportation research community to realize there are empirical traffic phenomena that call into question basically all fundamentals of the standard traffic flow theories and models as well as to accept the three-phase traffic theory solving the problem.

In this volume of the *Encyclopedia of Complexity and Systems Science*, the three-phase traffic theory and standard traffic flow theories as well as their applications for traffic management are discussed in several review articles written by researchers of different scientific groups. In addition to the consideration of the complexity of vehicular traffic, the volume includes reviews about pedestrian traffic research and evacuation phenomena as well as air traffic management.

Acknowledgments

I would like to thank Robert Meyers, the Editor-in-Chief of the *Encyclopedia of Complexity and Systems Science*, who proposed the project of this volume to me and then helped me very much in the project realization. I would like to thank Meghna Singh, Neha Thapa, Vignesh M., Hinduja Dhanasegaran, Andrew Spencer, and David Packer of the Springer Team for their outstanding editorial efforts in producing this volume of the encyclopedia.

I would like to thank all authors of this volume of the encyclopedia for discussions and fruitful cooperation. I thank all referees of the articles included in the volume whose critical comments have helped the authors to make their articles more readable and understandable.

I thank Michael Schreckenberg, Hubert Rehborn, Sergey Klenov, and Craig Davis for their support, help, and advice.

I thank our partners for their support in the project “MEC-View – Object detection for autonomous driving based on Mobile Edge Computing,” funded by the German Federal Ministry for Economic Affairs and Energy.

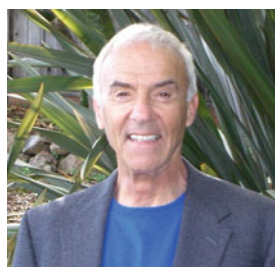
Finally, I thank my wife, Tatiana Kerner, for her help and understanding.

Contents

Complex Dynamics of Traffic Management: Introduction	1
Boris S. Kerner	
Breakdown in Traffic Networks	21
Boris S. Kerner	
Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks	79
Hesham Rakha and Aly Tawfik	
Optimization and Control of Urban Traffic Networks	131
Nathan H. Gartner and Chronis Stamatiadis	
Freeway Traffic Management and Control	167
Andreas Hegyi, Tom Bellemans and Bart De Schutter	
Modeling Approaches to Traffic Breakdown	195
Boris S. Kerner	
Mathematical Probabilistic Approaches to Traffic Breakdown ...	285
Boris S. Kerner and Sergey L. Klenov	
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory	313
Junfang Tian, Chenqiang Zhu and Rui Jiang	
Autonomous Driving in the Framework of Three-Phase Traffic Theory	343
Boris S. Kerner	
Spatiotemporal Features of Traffic Congestion	387
Boris S. Kerner	
Traffic Prediction of Congested Patterns	501
H. Rehborn, Sergey L. Klenov and M. Koller	
Physics of Mind and Car-Following Problem	559
Ihor Lubashevsky and Kaito Morimura	
Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems	593
Takashi Nagatani	

Travel Behavior and Demand Analysis and Prediction	613
Konstadinos G. Goulias	
Modelling of Pedestrian and Evacuation Dynamics	649
Mohcine Chraïbi, Antoine Tordeux, Andreas Schadschneider and Armin Seyfried	
Empirical Results of Pedestrian and Evacuation Dynamics	671
Maik Boltes, Jun Zhang, Antoine Tordeux, Andreas Schadschneider and Armin Seyfried	
Natural Disasters and Evacuations as a Communication and Social Phenomenon	701
Douglas Goudie	
Complex Dynamics of Air Traffic Flow	741
Banavar Sridhar and Kapil Sheth	
Index	753

About the Editor-in-Chief



Dr. Robert A. Meyers

President: RAMTECH Limited

Manger, Chemical Process Technology, TRW Inc.

Post doctoral Fellow: California Institute of Technology

Ph.D. Chemistry, University of California at Los Angeles

B.A. Chemistry, California State University, San Diego

Biography

Dr. Meyers has worked with more than 20 Nobel laureates during his career and is the originator and serves as Editor-in-Chief of both the Springer Nature *Encyclopedia of Sustainability Science and Technology* and the related and supportive Springer Nature *Encyclopedia of Complexity and Systems Science*.

Education

Postdoctoral Fellow: California Institute of Technology

Ph.D. in Organic Chemistry, University of California at Los Angeles

B.A. Chemistry with minor in Mathematics, California State University, San Diego

Dr. Meyers holds more than 20 patents and is the author or Editor-in-Chief of 12 technical books including the *Handbook of Chemical Production Processes*, *Handbook of Synfuels Technology*, and *Handbook of Petroleum Refining Processes* now in 4th Edition, and the *Handbook of Petrochemical*

Production Processes, now in its second edition (McGraw-Hill), and the *Handbook of Energy Technology and Economics*, published by John Wiley & Sons; *Coal Structure*, published by Academic Press; and *Coal Desulfurization* as well as the *Coal Handbook* published by Marcel Dekker. He served as Chairman of the Advisory Board for *A Guide to Nuclear Power Technology*, published by John Wiley & Sons, which won the Association of American Publishers Award as the best book in technology and engineering.

About the Volume Editor



Boris S. Kerner is the pioneer of the *three-phase traffic theory* and the *breakdown minimization principle*.

Biography

Boris Kerner was born in Moscow in 1947 and graduated from the Moscow Technical University MIREA in 1972. Boris Kerner received his Ph.D. and Sc. D. (Doctor of Sciences) from the Academy of Sciences of the Soviet Union, respectively, in 1979 and 1986. Between 1972 and 1992, his major interests include the physics of semiconductors, plasma, and solid-state physics as well as the development of a theory of autosolitons – solitary intrinsic states, which form in a broad class of physical, chemical, and biological dissipative systems.

Since 1992 Boris Kerner worked for the Daimler Company in Stuttgart, Germany. His major interest since then was the understanding of vehicular traffic. The empirical nucleation nature of traffic breakdown at highway understood by Boris Kerner is the basis for three phase traffic theory, which he introduced and developed in 1996–2002. Between 2000 and 2013, Boris Kerner was Head of a scientific research field *Traffic* at the Daimler Company in Stuttgart (Germany). In 2011, Boris Kerner was awarded with the degree *Professor* at the University of Duisburg-Essen in Germany. After his retirement from the Daimler Company on January 31, 2013, Prof. Kerner works at the University of Duisburg-Essen.

Boris Kerner is the author of more than 250 scientific works and patents as well as four books that are devoted to a variety of physical systems, complex dynamics of traffic flow in traffic and transportation networks, applications of

diverse intelligent transportation systems for traffic prognosis, traffic control, dynamic traffic assignment, as well as a study of autonomous and connected vehicles in mixed traffic flow.

Contributions to Traffic and Transportation Science

The three-phase traffic theory introduced by Kerner at the end of the 1990s is the framework for understanding of states of empirical traffic flow in three phases: (1) free flow (F), (2) synchronized flow (S), and (3) wide moving jam (J). The synchronized flow and wide moving jam traffic phases belong to congested traffic.

The main reason for the three-phase traffic theory is the explanation of the empirical nucleation nature of a transition from free flow to congested traffic called traffic breakdown. In the three-phase theory, it is assumed that traffic breakdown is a random (probabilistic) phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) that occurs in a metastable state of free flow. The main prediction of the three-phase theory is that this metastability of free flow with respect to the $F \rightarrow S$ transition is governed by the nucleation nature of an $S \rightarrow F$ instability that development leads to a local phase transition from synchronized flow to free flow ($S \rightarrow F$ transition).

The basic result of the three-phase theory about the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) shows that the three-phase theory is incommensurable with all earlier traffic flow theories. In 2011–2014, Boris Kerner expanded the three-phase theory, which he developed initially for highway traffic, for the description of city traffic.

The breakdown minimization (BM) principle introduced by Kerner in 2011 is devoted to management, control, and optimization of traffic and transportation networks. The BM principle permits to maximize the network throughput preventing traffic breakdown in the network.

The three-phase traffic theory and the BM principle are applied for the development of reliable methods of management, control, dynamic traffic assignment, and optimization of traffic and transportation networks and many other applications in traffic and transportation engineering, for example, for spatiotemporal reconstruction of empirical congested traffic patterns, for the development of autonomous driving vehicles that learn from human driving behavior, traffic flow simulation models, a reliable evaluation of intelligent transportation systems like studies of the effect of autonomous driving vehicles, vehicle-to-vehicle communication, other cooperative driving systems on mixed traffic flow, etc.

Contributors

Tom Bellemans Hasselt University, Diepenbeek, Belgium

Maik Boltes Institute for Advanced Simulation, Forschungszentrum Jülich, Jülich, Germany

Mohcine Chraïbi Institute for Advanced Simulation, Forschungszentrum Jülich GmbH, Jülich, Germany

Bart De Schutter Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

Nathan H. Gartner University of Massachusetts, Lowell, MA, USA

Douglas Goudie School of Earth and Environmental Sciences, James Cook University, Australian Centre for Disaster Studies, Townsville, QLD, Australia

Konstadinos G. Goulias University of California Santa Barbara, Santa Barbara, CA, USA

Andreas Hegyi Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands

Rui Jiang MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing, China

Boris S. Kerner Physics of Transport and Traffic, University Duisburg-Essen, Duisburg, Germany

Sergey L. Klenov Department of Physics, Moscow Institute of Physics and Technology, Moscow, Russia

M. Koller Research and Development RD/USN, HPC: 059-X901, Daimler AG, Sindelfingen, Germany

Ihor Lubashevsky University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, Japan

Kaito Morimura University of Aizu, Tsuruga, Ikki-machi, Aizu-Wakamatsu, Fukushima, Japan

Takashi Nagatani Department of Mechanical Engineering, Shizuoka University, Hamamatsu, Japan

Hesham Rakha Center for Sustainable Mobility, Virginia Tech, Transportation Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

H. Rehborn Research and Development RD/USN, HPC: 059-X901, Daimler AG, Sindelfingen, Germany

Andreas Schadschneider Institut für Theoretische Physik, Universität zu Köln, Köln, Germany

Armin Seyfried Institute for Advanced Simulation, Forschungszentrum Jülich GmbH, Jülich, Germany

School of Architecture and Civil Engineering, University of Wuppertal, Wuppertal, Germany

Kapil Sheth NASA Ames Research Center, Moffett Field, CA, USA

Banavar Sridhar NASA Ames Research Center, Moffett Field, CA, USA

Chronis Stamatiadis University of Massachusetts, Lowell, MA, USA

Aly Tawfik Center for Sustainable Mobility, Virginia Tech, Transportation Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

Junfang Tian Institute of Systems Engineering, College of Management and Economics, Tianjin University, Tianjin, China

Antoine Tordeux School of Mechanical Engineering and Safety Engineering, University of Wuppertal, Wuppertal, Germany

Jun Zhang State Key Laboratory of Fire Science, University of Science and Technology of China, Hefei, China

Chenqiang Zhu Institute of Systems Engineering, College of Management and Economics, Tianjin University, Tianjin, China

Complex Dynamics of Traffic Management: Introduction

Boris S. Kerner

Physics of Transport and Traffic, University
Duisburg-Essen, Duisburg, Germany

Article Outline

Glossary

Importance of Traffic Management

Standard Traffic and Transportation Theories

Paradigm Shift in Traffic and Transportation
Science

Objective

Contributions

Bibliography

Glossary

Air traffic flow Air traffic flow represents the distribution of air traffic over a region of space. Air traffic is undergoing major changes both in developed and developing countries. The demand for air traffic depends on population growth and other economic factors. Air traffic in the United States is expected to grow to 2 or 3 times the baseline levels of traffic in the next few decades. An understanding of the characteristics of the baseline and future flows is essential to design a good traffic flow management strategy.

Autonomous driving An autonomous driving vehicle is a self-driving vehicle that can move without a driver. Autonomous driving is realized through the use of an automated system in a vehicle: The automated system has control over the vehicle in traffic flow. Autonomous driving vehicle is also called self-driving, automated driving, or automatic driving vehicle.

Autonomous driving in framework of three-phase traffic theory An autonomous driving in the framework of the three-phase traffic

theory is the autonomous driving for which there is no fixed time headway to the preceding vehicle. This means the existence of an indifference zone in car-following for the autonomous driving vehicle.

Bottleneck Traffic breakdown leading to the onset of congestion in a traffic network occurs usually at a network bottleneck. Road bottlenecks are caused, for example, by on- and off-ramps, road gradients, roadworks, a decrease in the number of road lines (in the flow direction), traffic signal, etc.

Breakdown minimization (BM) principle The BM principle states that the optimum of a traffic or transportation network with N bottlenecks is reached when dynamic traffic assignment, optimization, and control are performed in the network in such a way that the probability for the occurrence of traffic breakdown in at least one of the network bottlenecks during a given time interval for observing traffic flow reaches the minimum possible value. The BM principle is the theoretical fundamental of transportation science that permits to maximize the network throughput preventing traffic breakdown in the network.

Cellular automata (CA) traffic flow model Cellular automata (CA) traffic flow model belongs to a class of microscopic traffic flow models. In the CA model, the time and space are discrete, and the evolution is described by update rules.

Community A group of neighbors or people with a commonality of association and generally defined by location, shared experience, or function.

Congested pattern control approach Congested pattern control approach is consistent with the empirical nucleation nature of traffic breakdown. In congested pattern control approach, no control of traffic flow at a network bottleneck is realized as long as free flow is realized at the bottleneck. Only after traffic breakdown has occurred, traffic control starts. Due to the congested pattern control approach, either free

flow recovers at the bottleneck or traffic congestion is localized at the bottleneck.

Crowd A large group of pedestrians who have gathered together.

Disaster The interface between an extreme physical event and a vulnerable human population.

Dynamic traffic assignment Dynamic traffic assignment is traffic assignment in a traffic or transportation network considering the temporal dimension of the assignment problem.

Empirical fundamental of transportation science The nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck is an empirical fundamental of transportation science.

Evacuation People relocating to safety escape hazardous disaster impacts. To move from a high danger zone to relative safety.

Incommensurability of standard traffic theories with three-phase traffic theory Three-phase traffic theory has been introduced for the explanation of metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck observed in real field traffic data. This metastability of free flow has no sense for classical (standard) traffic and transportation theories that are related to the state-of-the-art in transportation research. In accordance with the classical theory by Kuhn about scientific revolutions, this shows the incommensurability of the three-phase theory and the standard traffic flow theories. The term “incommensurable” has been introduced by Kuhn to explain a paradigm shift in the scientific field.

Intelligent transportation system (ITS) An intelligent transportation system (ITS) is an application which aims to provide innovative services relating to different modes of traffic management and enable users to be better informed and make safer, more coordinated, and “smarter” use of transport networks.

Measurements of traffic flow variables Macroscopic traffic variables like flow rate, vehicle speed, and occupancy are measured with local detectors in a traffic network. Probe vehicle data (called also floating car data (FCD)) includes vehicle positions as time functions measured by probe vehicles in traffic flow.

Mixed traffic flow Mixed traffic flow consists of autonomous driving vehicles that are randomly distributed between human driving vehicles.

Network throughput maximization approach A network throughput maximization approach is an approach to dynamic traffic assignment and control in a traffic or transportation network that maximizes the network throughput while keeping free flow conditions in the whole network.

Nonlinear map model A nonlinear map is a recurrence relation presented by a nonlinear function. The nonlinear map model is a dynamic model of bus motion described by the nonlinear map.

Optimal control Optimal control is a control methodology that formulates a control problem in terms of a performance function, also called an objective function. This function expresses the performance of the system over a given period of time, and the goal of the controller is to find the control signals that result in optimal performance.

Paradigm shift in traffic and transportation science The paradigm shift in traffic and transportation science is the fundamental change in the meaning of stochastic highway capacity because the meaning of highway capacity is the basis for the development of any method for traffic control, management, and organization of a traffic network as well as ITS-applications. The paradigm of standard traffic and transportation theories is that at any time instant there is a value of stochastic highway capacity. When the flow rate at a bottleneck exceeds the capacity value at this time instant, traffic breakdown must occur at the bottleneck. The new paradigm of traffic and transportation science following from the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) and the three-phase traffic theory changes fundamentally the meaning of stochastic highway capacity as follows. At any time instant, there is a range of highway capacity values between a minimum and a maximum highway capacity, which are themselves stochastic values. When the flow rate at a bottleneck is inside this capacity range related to this time instant, traffic breakdown can occur at the

bottleneck only with some probability, i.e., in some cases traffic breakdown occurs; in other cases it does not occur.

Pedestrian A pedestrian is a person traveling on foot.

Prediction of spatiotemporal congested traffic pattern The function of the prediction approach is to forecast the whole congestion pattern in space and time. Parameters of congested patterns at specific bottlenecks can provide a “fingerprint” of the road network, because they are typical, recurring, and specific to the related bottleneck. These features facilitate the predictability of traffic congestion.

Probe vehicle data The GPS (Global Positioning System) in a vehicle offers a position accuracy in the order of 5–10 m. Connected vehicles and several mobile devices (in vehicles) send their driving positions in small time intervals to central servers: the reconstruction of the traffic state is now based on masses of such probe vehicle data. The probe vehicle data gives the opportunity to retrieve the current traffic situation on all roads of a network.

Physics of mind Physics of mind is a branch of physics studying mental phenomena based on principles of the mind and their mathematical formulations as well as the properties and dynamics of systems governed by humans. It turns to mechanisms of emotions, cognition, concepts, intuitions, imagination, etc., as well as takes into account human intentionality, goal-oriented behavior, volition, memory effects, the bounded capacity of human cognition in the framework of phenomenology.

Spatiotemporal congested traffic pattern Spatiotemporal congested traffic patterns are traffic patterns in a traffic network that emerge and develop in space and time as a result of traffic breakdown in the traffic network. Traffic breakdown occurs usually at network bottlenecks.

Standard traffic and transportation theories We use the term “standard” for traffic and transportation theories, traffic flow modeling approaches, as well as methods for traffic routing, dynamic traffic assignment, and traffic control that are related to the state-of-the-art in the traffic and transportation scientific community.

Three-phase traffic theory Three-phase traffic theory is the framework for the understanding of empirical states of vehicular traffic flow in three phases: (i) free flow (F), (ii) synchronized flow (S), (iii) wide moving jam (J). The synchronized flow and wide moving jam phases belong to congested traffic. The three-phase traffic theory is the theoretical fundamental of transportation science that explains the empirical nucleation nature of traffic breakdown.

Traffic assignment Traffic assignment is a procedure used to find the network link flow rates from the original-destination demand.

Traffic breakdown Traffic breakdown is a transition from free vehicular traffic flow (free flow for short) to congested traffic in a traffic network. Traffic breakdown occurs usually at network bottlenecks. In empirical observations, traffic breakdown at a highway bottleneck occurs in a metastable state of free flow with respect to a transition from free flow to synchronized flow ($F \rightarrow S$ transition). Synchronized flow is one of two phases of congested traffic. Empirical traffic breakdown ($F \rightarrow S$ transition) at the bottleneck can be either spontaneous or induced.

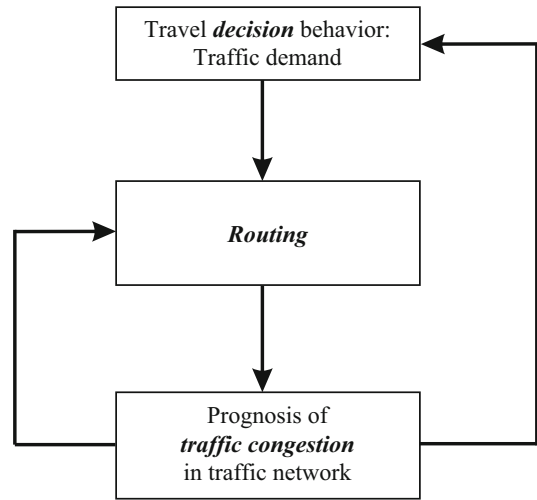
Traffic flow model A vehicular traffic flow model is devoted to the explanation and simulation of traffic flow phenomena, which are observed in measured data of real traffic flow, and to the prediction of new traffic flow phenomena that could be found in real traffic flow. First of all, a traffic flow model should explain and predict empirical (measured) features of traffic breakdown.

Traffic routing Traffic routing is a procedure that computes the sequence of roadways that vehicles should use moving in a traffic network. The procedure of standard traffic routing minimizes some utility objective function that could either be travel time or a generalized function that also includes road tolls. Traffic routing based on the network throughput maximization approach maximizes the network throughput while keeping free flow conditions in the whole network.

Traffic signal control Traffic control signals are defined as power-operated traffic devices

which alternately direct traffic to stop and to proceed. More specifically, traffic signals are used to control the assignment of right-of-way at locations where conflicts exist or where passive devices, such as signs and markings, do not provide the necessary flexibility of control to properly move traffic in a safe and efficient manner.

Travel demand The amount of travel within a time interval such as number of trips in a day, total amount of distance and total amount of travel time, the locations (destinations) visited, the means used to reach these locations, departure time and arrival time of trips, routes followed in reaching these locations, the sequencing and assembly of trips in groups, and the purpose or activity engaged in at the end of each trip.



Complex Dynamics of Traffic Management: Introduction, Fig. 1 Explanation of complexity of vehicular traffic in a network

Importance of Traffic Management

Traffic management is an extremely complex dynamic process associated with the spatiotemporal behavior of many-particle nonlinear systems. Between review articles related to *complex dynamics of traffic management* of the Encyclopedia of Complexity and Systems Science, there are several review articles devoted to various aspects of the complex dynamics of vehicular traffic management (Gartner and Stamatiadis 2009; Goulias 2009; Hegyi et al. 2017; Kerner 2017c, 2018a, b, c; Kerner and Klenov 2018; Lubashevsky and Morimura 2018; Nagatani 2017; Rakha and Tawfik 2009; Rehborn et al. 2017; Tian et al. 2018), air traffic management (Sridhar and Sheth 2019), and pedestrian traffic management (Boltes et al. 2018; Chraibi et al. 2018; Goudie 2017).

In particular, the complexity of vehicular traffic in a traffic network is due to nonlinear interactions between the following three main dynamic processes (Fig. 1):

- (i) Travel *decision* behavior, which determines traffic demand in a traffic network (Goulias 2009)
- (ii) *Routing* of vehicles in the network (Kerner 2017c; Rakha and Tawfik 2009)
- (iii) *Prognosis of traffic congestion* within the network (Kerner 2017c, 2018a; Rehborn et al. 2017).

Travel decision behavior determines travel demand. Traffic routing in the network is associated with traffic supply. However, traffic congestion occurring within the traffic network restricts free flow travel. This influences both travel decision behavior and traffic routing in the network. Indeed, because of traffic congestion, a person decides to stay at home or travel by train rather than by car. A feedback between *prognosis of traffic congestion* and *travel decision* is symbolically shown by arrow on the right hand side in Fig. 1. In turn, because of traffic congestion on a route from an origin to a destination usually used, a person changes the route of travel. A feedback between *prognosis of traffic congestion* and *routing* is symbolically shown by arrow on the left hand side in Fig. 1.

Routing based on a traffic optimization together with the prognosis of traffic congestion is called *dynamic traffic assignment* in the traffic network. A dynamic traffic assignment model should find time-functions of the link inflows for the traffic network. The model includes usually a traffic flow

model, which makes a prognosis of traffic in the network, and a routing model associated with the problem of traffic optimization. The traffic flow and router models are connected by a feedback loop (see the feedback loop between *prognosis of traffic congestion* and *routing* in Fig. 1). As a result, traffic congestion in the network predicted by the traffic flow model changes results of dynamic traffic assignment considerably. For this reason, the traffic flow model should make the prediction of the occurrence of traffic congestion as close as possible to real features of traffic breakdown and resulting traffic congestion found in empirical observations.

The necessity of vehicular traffic management is associated with the occurrence of traffic congestion in a traffic network at a large enough traffic demand. The occurrence of traffic congestion results in a considerably increase in travel time, fuel consumption, and other travel costs as well as in a decrease in traffic safety and comfort. Therefore, through the management of vehicular traffic, the negative effects of traffic congestion should be decreased.

Vehicular traffic management can be realized through *dynamic traffic assignment* and *traffic control* in a traffic network. Methods of traffic control include on-ramp metering, variable speed limit control, traffic signal control at signalized road intersections, and diverse approaches to cooperative driving systems. Cooperative driving systems include, for example, vehicle assistance systems, automated driving (called also automatic driving, or self-driving, or else autonomous driving) vehicles, as well as systems that are based on the use of vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2X) communication and many other intelligent transportation systems (ITS) (Gartner and Stamatiadis 2009; Hegyi et al. 2017; Kerner 2017c, 2018c; Rakha and Tawfik 2009).

Vehicular traffic congestion results from *traffic breakdown*. Traffic breakdown is a transition from free flow to congested traffic in a traffic or transportation network. Traffic breakdown occurs usually at network bottlenecks. A bottleneck can be a result of road works, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, traffic signal, bad weather conditions, accidents, etc. Thus, to develop reliable

methods for dynamic traffic assignment and traffic control in traffic and transportation networks, empirical features of traffic breakdown should be understood.

Standard Traffic and Transportation Theories

We will use the term “standard” for traffic and transportation theories, traffic flow modeling approaches, as well as methods for dynamic traffic assignment and traffic control that are related to the state-of-the-art in the traffic and transportation scientific community.

First approaches to traffic and transportation science appeared in 1920s–1930s. One of the most classical works is widely considered the work by Greenshields et al. (1935) about a study of traffic capacity. Fundamentals of standard vehicular traffic and transportation science that led to the most of the approaches for traffic simulation, control, and dynamic traffic assignment, which are related to the state-of-the-art in standard traffic and transportation research, were developed in 1950s–1960s.

Lighthill and Whitham (1995) and Richards (1956) developed the famous Lighthill-Whitham-Richards (LWR) macroscopic traffic flow model. A gas-kinetic traffic flow model was introduced by Prigogine (1961).

First microscopic (car-following) traffic flow models were introduced by Reuschel (1950a, b), Pipes (1953), Kometani and Sasaki (1958, 1959, 1961a, b), and Herman, Gazis, Rothery, Montroll, Chandler, and Potts (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959) who introduced the well-known GM (General Motors) microscopic traffic flow model as well as by Newell (1961). Traffic flow instability studied in the GM model in Chandler et al. (1958), Gazis et al. (1959, 1961), and Herman et al. (1959) can be considered the classical traffic flow instability.

To predict vehicle delay characteristics at road bottlenecks, queueing models employing the methods of probability and statistics have been applied (see, e.g., Edie (1954), Moskowitz (1954) and references in the books (Gerlough and Huber 1975; Newell 1982)).

Webster (1958) introduced a model of traffic at traffic signal that is up to now the basis of the most standard approaches for traffic signal control.

Considering the result of the description of traffic breakdown at a bottleneck with traffic flow models independent of a diverse variety of mathematical model formulations, we can classified all standard traffic flow models into two different classes: (i) LWR model class and (ii) GM model class (with traffic flow instability) (see for a review Kerner (2018a)).

Wardrop (1952) introduced user equilibrium (UE) and system optimum (SO) equilibrium principles. The most of the standard approaches to dynamic traffic assignment that are devoted to the minimization of network travel times (or other travel costs) are based on the use of either Wardrop's UE or Wardrop's SO or else their combinations (see, e.g., reviews and books (Bell and Iida 1997; Chiu et al. 2011; Friesz and Bernstein 2016; Kachroo and Özbay 2018; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984)).

The mentioned classical works (Chandler et al. 1958; Gazis et al. 1959, 1961; Greenshields et al. 1935; Herman et al. 1959; Kometani and Sasaki 1958, 1959, 1961a, b; Lighthill and Whitham 1995; Newell 1961; Pipes 1953; Prigogine 1961; Reuschel 1950a, b; Richards 1956; Wardrop 1952; Webster 1958) are the fundamentals for standard traffic flow theories and models, standard approaches to dynamic traffic assignment and traffic control (e.g., reviews and books (Ashton 1966; Bell and Iida 1997; Bellomo et al. 2002; Brackstone and McDonald 1999; Brockfeld et al. 2003; Buslaev et al. 2013; Chiu et al. 2011; Chowdhury et al. 2000; Daganzo 1997; Drew 1968; Elefteriadou 2014; Ferrara et al. 2018; Friesz and Bernstein 2016; Gartner et al. 1997, 2001; Gartner and Stamatidis 2009; Gasnikov et al. 2013; Gerlough and Huber 1975; Hegyi et al. 2017; Helbing 2001; Highway Capacity Manual 2000, 2016; Kachroo and Özbay 2018; Kessels 2019; Leutzbach 1988; Maerivoet and De Moor 2005; Mahmassani 2001; Mahnke et al. 2009; Mannering and Kilareski 1998; May 1990; Nagatani 2002; Nagel et al. 2003; Newell 1963, 1982; Papageorgiou and Papamichail 2008; Peeta and Ziliaskopoulos 2001; Piccoli and Tosin 2009; Prigogine and Herman 1971; Rakha

and Tawfik 2009; Ran and Boyce 1996; Roess and Prassas 2014; Saifuzzaman and Zheng 2014; Seo et al. 2017; Schadschneider et al. 2011; Sheffi 1984; Shvetsov 2003; Treiber and Kesting 2013)). The standard traffic flow models and traffic control methods and approaches to dynamic traffic assignment are related to the state-of-the-art in traffic and transportation research.

Based on these standard traffic flow models and approaches to dynamic traffic assignment, a number of traffic and transport simulation tools have been developed like TRANSIMS (Transportation ANalysis and SIMulation System) (TRANSIMS Open Source 2013; TRANSIMS 2017) MATSim (Multi-Agent Transport Simulation) (Horni et al. 2016) and other well-known simulation tools VISSIM, AVENUE, Paramics, Aimsun, MITSIMLab, SUMO, DRACULA, Dynameq, DynaMIT, METANET reviewed in the book by Barceló (2010).

The standard traffic flow theories and models had a great impact on the understanding of many empirical traffic phenomena. However, users of traffic and transportation networks would expect that through the use of traffic control, dynamic traffic assignment and other methods of dynamic optimization traffic breakdown can be prevented, i.e., free flow can be maintained in the network. This is because as above-mentioned due to traffic breakdown congested traffic occurs in which travel time, fuel consumption as well as other travel costs increased considerably in comparison with travel costs in free flow.

- Therefore, any traffic and transportation theory applied for the development of autonomous driving vehicles and reliable methods of dynamic traffic assignment and traffic control should be consistent with the empirical features of traffic breakdown at network bottlenecks.

Paradigm Shift in Traffic and Transportation Science

Empirical Fundamental of Transportation Science

The most important empirical feature of traffic breakdown in a network is the nucleation nature of traffic breakdown found in real field traffic data.

The term *nucleation nature* of traffic breakdown means that traffic breakdown at a network bottleneck occurs in a metastable free flow. The metastability of free flow is as follows. There can be many speed (density, flow rate) disturbances in free flow. Amplitudes of the disturbances can be very different. When a disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown occurs. Such a disturbance resulting in the breakdown is called *nucleus* for the breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays; as a result, no traffic breakdown occurs.

- The empirical nucleation nature of traffic breakdown at a network bottleneck is an empirical fundamental of transportation science.
- For this reason, traffic and transportation theories, which are not consistent with the nucleation nature of traffic breakdown at highway bottlenecks, cannot be applied for the development of reliable traffic control, dynamic traffic assignment as well as other reliable applications of intelligent transportation systems (ITS) in traffic and transportation networks.

Failure of Standard Traffic Theories in Real World Applications

Unfortunately, standard traffic and transportation theories and traffic control methods mentioned in section “[Standard Traffic and Transportation Theories](#),” which have had a great impact on the understanding of many empirical traffic phenomena, have nevertheless failed by their applications in the real world. Even several decades of a very intensive effort to improve and validate standard network optimization and control models based on the classical traffic and transportation theories have had no success. Indeed, there can be found no examples where on-line implementations of the standard network optimization models based on these classical (standard) traffic and transportation theories and methodologies could reduce congestion in real traffic and transportation networks. A discussion of this critical statement of the author can be found in the book by Kerner (2017b) as well as in reviews of the Encyclopedia by Kerner (2017c, 2018a, b) in which a comparison of

empirical congestion data and the prediction of standard models and methods has been made.

Additionally, in the article of the Encyclopedia (Kerner 2017c), it will be shown that applications of the standard approaches for dynamic traffic assignment in traffic networks, which are related to the state-of-the-art in traffic and transportation research (see, e.g., reviews and books (Bell and Iida 1997; Chiu et al. 2011; Friesz and Bernstein 2016; Kachroo and Özbay 2018; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984)), deteriorate the traffic system while provoking heavy traffic congestion in urban networks.

- This failure of classical (standard) traffic and transportation theories for ITS-applications can be explained as follows: The standard traffic and transportation theory are not consistent with the empirical nucleation nature of traffic breakdown.

The nucleation nature of empirical traffic breakdown was understood by the author at the end of 1990s (first review of the empirical nucleation features of traffic breakdown at road bottlenecks was done in 2004 (Kerner 2004); in the Encyclopedia, these features are considered in (Kerner 2017c). In contrast, the generally accepted fundamentals and methodologies of traffic and transportation theory were introduced in the 1950s–1960s (section “[Standard Traffic and Transportation Theories](#)”). Thus, the scientists whose ideas led to the classical (standard) traffic and transportation theories could not know and, therefore, they did not appreciate the empirical nucleation nature of traffic breakdown (see (Chandler et al. 1958; Gazis et al. 1959, 1961; Greenshields et al. 1935; Herman et al. 1959; Kometani and Sasaki 1958 1959, 1961a, b; Lighthill and Whitham 1995; Newell 1961; Pipes 1953; Prigogine 1961; Reuschel 1950a, b; Richards 1956; Wardrop 1952; Webster 1958) and reviews (Ashton 1966; Bell and Iida 1997; Brackstone and McDonald 1999; Daganzo 1997; Drew 1968; Gartner et al. 1997; Gerlough and Huber 1975; Highway Capacity Manual 2000; Leutzbach 1988; Mannering and Kilareski 1998;

May 1990; Newell 1963, 1982; Prigogine and Herman 1971)). It appears that many traffic researchers do not accept the empirical nucleation nature of traffic breakdown today (see reviews and books (Chiu et al. 2011; Elefteriadou 2014; Ferrara et al. 2018; Gartner and Stamatiadis 2009; Highway Capacity Manual 2016; Kessels 2019; Papageorgiou and Papamichail 2008; Piccoli and Tosin 2009; Rakha and Tawfik 2009; Saifuzzaman and Zheng 2014; Seo et al. 2017; Treiber and Kesting 2013)).

Theoretical Fundamentals of Transportation Science

To explain the empirical nucleation nature of traffic breakdown, the author introduced the three-phase traffic theory (three-phase theory for short). The three-phase theory is the framework for understanding empirical states of traffic flow in three phases: (i) free flow (F), (ii) synchronized flow (S), and (iii) wide-moving jam (J). The synchronized flow and wide-moving jam phases belong to congested traffic (see reviews (Kerner 2004, 2017c)).

To maximize the network throughput while keeping free flow in a traffic network, the author introduced the breakdown minimization (BM) principle. The BM principle states that the optimum of a traffic or transportation network with N bottlenecks is reached, when dynamic traffic assignment, optimization, and/or control are performed in the network in such a way that the probability for the occurrence of traffic breakdown in at least one of the network bottlenecks during a given time interval for observing traffic flow reaches the minimum possible value. The BM principle is equivalent to the maximization of the probability that during the time interval for observing traffic flow, traffic breakdown occurs at

none of the network bottlenecks. The BM principle and its possible applications have been reviewed in Kerner (2017c).

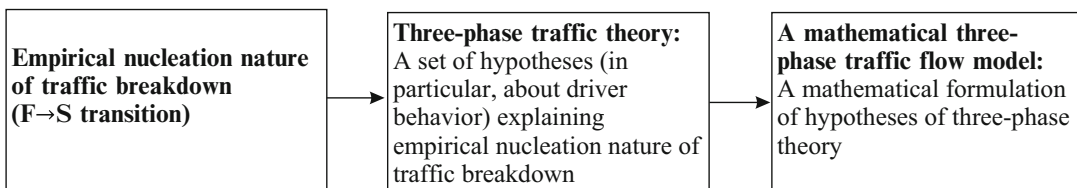
- The three-phase theory is the theoretical fundamental of transportation science that explains the empirical nucleation nature of traffic breakdown.
- The BM principle is the theoretical fundamental of transportation science that permits to maximize the network throughput preventing traffic breakdown in the network.

The Origin of Three-Phase Traffic Theory

Vehicular traffic in traffic and transportation networks occurs in space and time. Therefore, to understand vehicular traffic, real field traffic data measured in space and time should be understood. The most important empirical spatiotemporal traffic phenomenon is the nucleation nature of traffic breakdown at a network bottleneck. For this reason, the empirical nucleation nature of traffic breakdown is the origin of the three-phase theory (Fig. 2). The three-phase theory consists of a set of hypotheses that explain the empirical nucleation nature of traffic breakdown as well as features of resulting traffic congestion in a traffic network. There could be a diverse variety of mathematical formulations of the hypotheses of the three-phase theory.

To illustrate that terms and definitions used in the three-phase theory are solely based on the empirical nucleation nature of traffic breakdown as well as features of resulting traffic congestion in a traffic network (Fig. 2), we consider the definitions of the traffic phases S and J of congested traffic made in the three-phase theory (Fig. 3).

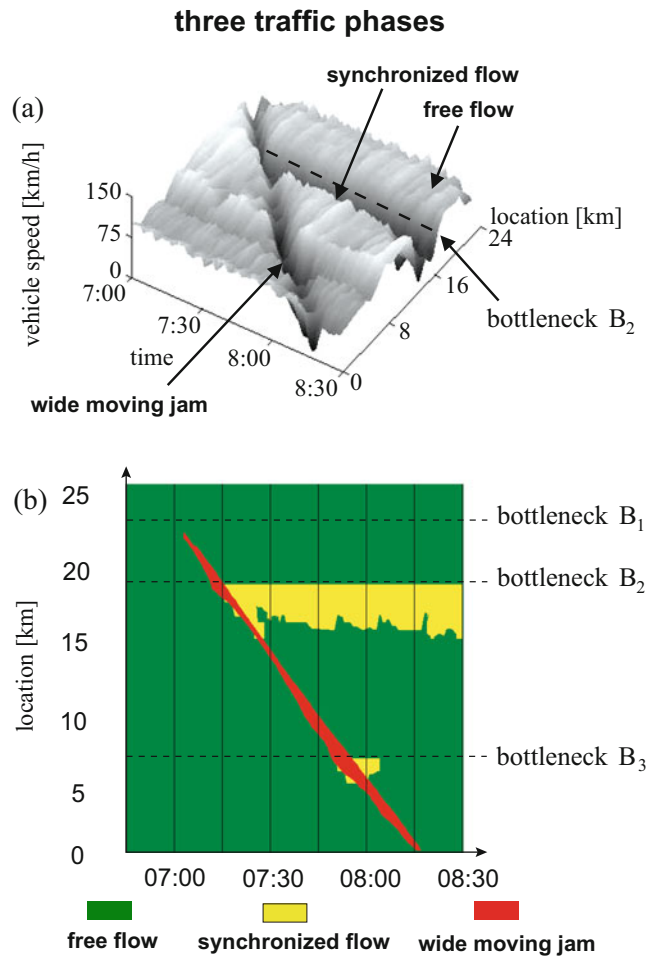
A congested traffic pattern (congested pattern for short) is a spatiotemporal traffic pattern within



Complex Dynamics of Traffic Management: Introduction, Fig. 2 Explanation of the origin of the three-phase theory

Complex Dynamics of Traffic Management: Introduction,

Fig. 3 Explanation of empirical traffic phase definitions used in three-phase theory: (a) speed data in space and time (1 min averaged data) measured on a German freeway (a) and the same data presented in space and time through the use of some averaging method (b). (Adapted from Kerner (2004))



which there is congested traffic. A front of a congested pattern is either a moving or motionless region within which one or several of the traffic variables (speed, density, flow rate, etc.) change abruptly.

In Fig. 3, we can see that there are two qualitative different congested patterns: The downstream front of one of the patterns is fixed at an on-ramp bottleneck (bottleneck B₂ in Fig. 3). We call the road location in a vicinity of the bottleneck at which the downstream of the congested pattern is fixed as an effective location of the bottleneck (bottleneck location for short). The congested pattern whose downstream front is fixed at the bottleneck has been called *synchronized flow* (S). Within the downstream front of synchronized flow, vehicles accelerate from synchronized flow to free flow.

Another congested traffic pattern propagates as a whole localized traffic pattern along the road. The density is very large, whereas the speed and flow rate are very low (as low as zero) within this pattern called a *moving traffic jam* (moving jam for short). The moving jam in Fig. 3 propagates through free flow and synchronized flow as well as through the bottlenecks (bottlenecks B₂ and B₃ in Fig. 3) while maintaining the mean velocity of the downstream front of the pattern. The moving jam that exhibits this characteristic feature has been called *wide moving jam* (J).

- The empirical nucleation nature of traffic breakdown is the origin of the three-phase theory.
- The origin of the definitions of the traffic phases S and J are spatiotemporal empirical features of measured traffic data only.

Main Prediction of Three-Phase Traffic Theory

Empirical traffic breakdown at a highway bottleneck is a phase transition from free flow (F) to synchronized flow (S) ($F \rightarrow S$ transition) at the bottleneck.

The main prediction of the three-phase theory is that free flow at the bottleneck can be in a metastable state with respect to an $F \rightarrow S$ transition. This free flow metastability at the bottleneck explains the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition). In its turns, the nucleation nature of the $F \rightarrow S$ transition is governed by the nucleation nature of an $S \rightarrow F$ instability in synchronized flow (see the Encyclopedia article (Kerner 2017c)).

- The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a bottleneck.

The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. To reach this goal, in congested traffic a new traffic phase called synchronized flow has been introduced. The basic feature of the synchronized flow traffic phase formulated in the three-phase theory leads to the nucleation nature of the $F \rightarrow S$ transition. In this sense, the synchronized flow traffic phase, which ensures the nucleation nature of the $F \rightarrow S$ transition at a highway bottleneck, and the three-phase traffic theory can be considered synonymous.

Incommensurability of Three-Phase Traffic Theory and Classical Traffic Theories

The metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck has no sense for the classical (standard) traffic and transportation theories. The three-phase theory has been introduced for the explanation of this free flow metastability.

In accordance with the classical book by Kuhn (2012), this shows the incommensurability of the three-phase theory and the classical traffic flow theories. Therefore, the three-phase theory is incommensurable with all classical traffic flow theories (Kerner 2017c). The term “incommensurable” has been introduced by Kuhn (2012) to explain a paradigm shift in a scientific field.

Meaning of Highway Capacity

The paradigm shift in traffic and transportation science is the fundamental change in the meaning of stochastic highway capacity. This is because the meaning of highway capacity is the basis for the development of any method for traffic control, management, and organization of a traffic network as well as ITS-applications.

The *paradigm* of standard traffic and transportation theories can be formulated as follows: At a given time instant, there is a value of highway capacity. This is also true when it is assumed that at any time instant there is a stochastic value of highway capacity. This means that when the flow rate at a bottleneck exceeds the capacity value at this time instant, traffic breakdown *must* occur at the bottleneck. This classical (standard) meaning of highway capacity is the basis for any standard method for traffic control, management, and organization of a traffic network as well as ITS-applications (see, for example, reviews and books (Ashton 1966; Ran and Boyce 1996; Drew 1968; Prigogine and Herman 1971; Huber 1975; Newell 1982; Sheffi 1984; May 1990; Bell and Iida 1997; Daganzo 1997; Gartner et al. 1997, 2001; Leutzbach 1988; Mannering and Kilareski 1998; Brackstone and McDonald 1999; Highway Capacity Manual 2000, 2016; Chowdhury et al. 2000; Helbing 2001; Peeta and Ziliaskopoulos 2001; Mahmassani 2001; Nagatani 2002; Bellomo et al. 2002; Nagel et al. 2003; Brockfeld et al. 2003; Papageorgiou and Papamichail 2008; Piccoli and Tosin 2009; Rakha and Tawfik 2009; Barceló 2010; Chiu et al. 2011; Schadschneider et al. 2011; Treiber and Kesting 2013; Elefteriadou 2014; Roess and Prassas 2014; Saifuzzaman and Zheng 2014; Friesz and Bernstein 2016; Seo et al. 2017; Gerlough and Hegyi et al. 2017; Kachroo and Özbay 2018; Ferrara et al. 2018; Kessels 2019)).

The new paradigm, which follows from the empirical nucleation nature of traffic breakdown and the three-phase theory (Kerner 2004, 2017b,c), can be formulated as follows: *At any time instant*, there is a *range* of the flow rates within which traffic breakdown ($F \rightarrow S$ transition) *can* occur with some probability, but traffic breakdown *does not* necessarily occur.

The new paradigm fundamentally changes the meaning of stochastic highway capacity as follows.

There is a range of highway capacity values between a minimum and a maximum highway capacity at any time instant. When the flow rate at a bottleneck is inside this capacity range related to this time instant, traffic breakdown ($F \rightarrow S$ transition) can occur at the bottleneck only with some probability, i.e., in some cases traffic breakdown occurs, in other cases traffic breakdown does not occur. The existence *at any time instant* of the infinite number of highway capacity values within the capacity range means that highway capacity is stochastic. Both the minimum and maximum highway capacities are also stochastic values.

This basic change in the meaning of highway capacity results from the empirical nucleation nature of traffic breakdown at a bottleneck. Three-phase theory is a theoretical framework that explains highway capacity as a range of the flow rates within which traffic breakdown can occur with some probability. To be consistent with the empirical nucleation nature of traffic breakdown, other methods (in comparison with the methods developed in accordance with the standard meaning of highway capacity) for traffic control and management in traffic networks as well as other ITS-applications should be developed.

Strict Belief in Classical Theories as Reason for Defective Analysis of Empirical Traffic Phenomena

Almost each of the traffic models can explain some real traffic phenomena, and each of the models exhibits a limited area of the applicability for the explanation of real traffic and transportation phenomena. This is also related to the classical (standard) traffic flow theories of section “[Standard Traffic and Transportation Theories](#)” that are able to explain some of empirical traffic flow phenomena. As we have mentioned above, the classical traffic flow theories failed for ITS-applications because these theories and models are not consistent with the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck.

In the most fields of science, new experimental and/or empirical phenomena initiate the development of theories and models of the phenomena. Unfortunately, it is often different in the field of standard traffic flow theories: Usually, results of

classical traffic flow theories and solutions of associated traffic flow models determine the methodologies of analyzing of empirical traffic data. This prevents the understanding of those new real traffic flow phenomena that cannot be explained by the classical (standard) theories.

- A strict belief in classical traffic theories in traffic and transportation community leads to a defective (and sometimes invalid) analysis of empirical traffic phenomena.
- In particular, classical traffic flow theories and models do not show the nucleation nature ($F \rightarrow S$ transition) at a highway bottleneck. This is probably the reason why the empirical evidence of the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck is simply ignored in the state-of-the-art of traffic and transportation research (see, e.g., reviews and books (Barceló 2010; Bellomo et al. 2002; Brockfeld et al. 2003; Chiu et al. 2011; Chowdhury et al. 2000; Elefteriadou 2014; Ferrara et al. 2018; Friesz and Bernstein 2016; Gartner et al. 2001; Gartner and Stamatiadis 2009; Hegyi et al. 2017; Helbing 2001; Highway Capacity Manual 2016; Horni et al. 2016; Kachroo and Özbay 2018; Kessels 2019; Maerivoet and De Moor 2005; Mahmassani 2001; Mahnke et al. 2009; Nagatani 2002; Nagel et al. 2003; Papageorgiou and Papamichail 2008; Peeta and Ziliaskopoulos 2001; Piccoli and Tosin 2009; Rakha and Tawfik 2009; Roess and Prassas 2014; Saifuzzaman and Zheng 2014; Seo et al. 2017; Schadschneider et al. 2011; TRANSIMS Open Source 2013; TRANSIMS 2017; Treiber and Kesting 2013) and references there).

In general, we can distinguish two different standard invalid methodologies of a study of empirical traffic flow phenomena:

- (i) All empirical traffic flow phenomena that contradict solutions of a traffic model appear to be ignored.
- (ii) Real empirical spatiotemporal traffic flow phenomena are averaged either in space and/or in time. As a result of this averaging, important empirical spatiotemporal traffic

flow phenomena cannot be found in the averaged data any more. Often this averaging of real measured data makes the resulting “empirical traffic flow phenomenon” in the agreement with theoretical conclusions made from a traffic model. In other words, all empirical features of real traffic phenomena, which cannot be explained by the model, are eliminated through the data averaging.

Examples of the defective and invalid analysis of empirical traffic data have been discussed in Sects. 10.3.7 and 10.3.10 of Kerner (2009) as well as in Sect. 4.11 of Kerner (2017b).

From the point of view of the author, the invalid methodologies of such a study of empirical traffic flow phenomena can be explained by a very long time interval between the development of the classical traffic theories made in 1950s–1960s and the understanding of the empirical nucleation nature of traffic breakdown at road bottlenecks made at the end of 1990s. During this time interval, several generations of traffic research studies developed a huge number of traffic flow models based on the classical approaches. Thus, in accordance with historical analysis of scientific revolutions in other fields of science made by Kuhn (2012), we may assume that it is very difficult for traffic researchers to see and accept that there is empirical traffic phenomenon that contradicts basically all fundamentals of the classical traffic flow theories and models. Some prominent examples of these invalid methodologies of empirical analyzes of traffic flow phenomena have been considered in the book by Kerner (2017b) as well as in the Encyclopedia articles by Kerner (2017c, 2018a).

Success of Three-Phase Traffic Theory in Real World Applications

As mentioned above, standard traffic and transportation theories of the world-wide traffic research community failed for reliable ITS-applications in the real world.

Contrarily, the success of the three-phase theory is associated with its reliable ITS-applications made by the Daimler Company and in traffic control centers in Germany. In particular, some

of the ITS-applications of the three-phase theory like models ASDA/FOTO that are working in real installations during last 20 years at a traffic control center in Germany are presented in the review article by Rehborn et al. (2017); an application of the three-phase theory for autonomous driving used by the Daimler Company is considered in Kerner (2018c). Recently through a detailed study of empirical data of probe vehicles (Rempe et al. 2016, 2017), colleagues by the BMW Company have confirmed the need of the use of the three-phase theory for the development of reliable ITS-applications.

Objective

The standard traffic and transportation theories that failed by their applications in the real world are currently the methodologies of teaching programs in most universities, the subject of publications in most transportation research journals and scientific conferences.

Consequently, the main objective of review articles in the fields of *complex dynamics of traffic management* of the Encyclopedia of Complexity and Systems Science is to give readers the opportunity to decide themselves about the criticism of classical (standard) traffic and transportation theories made by the author in the books and reviews (Kerner 2004, 2009, 2013, 2015, 2016, 2017a, b).

To reach this goal, in the Encyclopedia we have tried to present such review articles that cover almost the whole spectrum of classical (standard) traffic and transportation theories (Gartner and Stamatiadis 2009; Hegyi et al. 2017; Rakha and Tawfik 2009) and their criticisms (Kerner 2017c, 2018a) together with review articles about the three-phase traffic theory (Kerner 2017c, 2018a; Kerner and Klenov 2018; Tian et al. 2018), the BM principle (Kerner 2017c), as well as their ITS-applications (Kerner 2017c, 2018a, c; Rehborn et al. 2017). Additionally, reviews about other important subjects of vehicular traffic (Goulias 2009; Lubashevsky and Morimura 2018; Nagatani 2017), air traffic management (Sridhar and Sheth 2019), and pedestrian traffic management (Boltes et al. 2018; Chraïbi et al. 2018; Goudie 2017) have been presented.

Contributions

Explanations of the empirical nucleation nature of traffic breakdown with the three-phase theory can be found in the article by Kerner (2017c). In this article, readers can find:

- (i) The empirical proof of the nucleation nature of traffic breakdown at highway bottlenecks
- (ii) Discussions of the main prediction of the three-phase theory as well as requirements for three-phase traffic flow models
- (iii) A method for traffic control called “congested pattern control approach” that is consistent with the empirical nucleation nature of traffic breakdown
- (iv) A critical discussion of the standard (classical) definition of stochastic highway capacity
- (v) A critical discussion of standard dynamic traffic assignment explaining that and why standard dynamic traffic assignment deteriorates the traffic system while provoking heavy traffic congestion in urban networks
- (vi) Criticisms of applications of standard traffic flow modeling approaches for on-ramp metering and for a study of the effect of autonomous (automated) driving vehicles on traffic flow
- (vii) The nucleation nature of traffic breakdown at traffic signal as well as
- (viii) Dynamic traffic assignment in a traffic network based on applications of the BM principle. It will be shown that one of the application of the BM principle is the maximization of the network throughput while maintaining free flow conditions in the whole network.

It must be emphasized that the criticisms of standard traffic and transportation theories and methodologies made in the article by Kerner (2017c) are *not* related to pioneer ideas of traffic management in traffic and transportation research, for example, through the use of on-ramp metering, speed limit control, dynamic traffic assignment, cooperative driving systems, autonomous driving,

etc. Rather than these pioneer ideas for traffic management, the criticism of the author made in the article by Kerner (2017c) is related to *standard models and methods* developed for the realization of these ideas: In the standard models and methods for the realization of the ideas for traffic management, the empirical nucleation nature of traffic breakdown (empirical fundamental of transportation science) is ignored. A direct comparison of using the standard models and methods to using three-phase theory has been made in reviews (Kerner 2017c, 2018a).

The standard methods for dynamic traffic assignment in traffic networks criticized in Kerner (2017c) are reviewed in the articles by Rakha and Tawfik (2009). Later developments of standard dynamic traffic assignment (see, e.g., for reviews (Barceló 2010; Chiu et al. 2011; Kachroo and Özbay 2018)) do not change qualitatively basic model approaches already reviewed in Rakha and Tawfik (2009): The approaches exhibit the same feature of the deterioration of the traffic system as that criticized in Kerner (2017c). Standard models for traffic control have been reviewed by Hegyi et al. (2017). A criticism of the standard models for traffic control has been made in Kerner (2017c, 2018a). Readers have a possibility to make their own impression about the validity of this criticism on the associated standard traffic and transportation theories and their ITS-applications reviewed in Hegyi et al. (2017) and Rakha and Tawfik (2009).

The review article of Goulias (2009) is devoted to a detailed consideration of the theoretical and modeling approaches to travel behavior analysis. As shown in Goulias’s article, traveler values and attitudes are referred to motivational, cognitive, situational, and disposition factors determining human behavior. Goulias presents a sequential chronological order merging technological innovations and theoretical innovations to the modeling of travel behavior. There are a diverse variety of approaches to the modeling of travel behavior, which are based on mathematical methods like multiagent microsimulation widely used in many other fields of systems science. Inputs to these models are the typical regional model data about social, economic, and demographic information

of potential travelers and land use information to create schedules followed by people in their everyday life. The output is detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity. The in-depth understanding of transportation-related human behavior is essential to all kinds of traffic management. This is because travel behavior refers primarily to the modeling and analysis of travel demand. Later developments in the field of travel behavior are devoted to the integration of the ideas of travel behavior models (Goulias 2009) (see also later reviews of travel behavior models in Castiglione et al. (2015), and Horni et al. (2016)) for the development of transport simulation tools in which both travel behavior, traffic flow, and dynamic traffic assignment and traffic control in traffic and transportation networks should be simulated. Examples of such tools have been considered in Barceló (2010), Horni et al. (2016), TRANSIMS Open Source (2013), and TRANSIMS (2017). However, as explained in Kerner (2017c, 2018a), standard models for traffic flow, dynamic traffic assignment, and traffic control simulations used in the simulation tools (see reviews of some of these tools in Barceló (2010), Horni et al. (2016), TRANSIMS Open Source (2013), and TRANSIMS (2017)) cannot be used for a reliable evaluation of ITS-applications (see also section “Standard Traffic and Transportation Theories”).

Gartner and Stamatiadis (2009) consider methods for standard traffic modeling and traffic signal control in city and urban traffic networks. In an urban network with short enough network links, dynamic traffic phenomena are determined mostly by traffic signals and other traffic regulations at link intersections. The dynamic traffic phenomena in the urban network can be nevertheless very complex, because traffic regulations at one of the link intersections can have a great influence on the probability of traffic congestion occurring on other links of the network. In their article, Gartner and Stamatiadis (2009) present a historical model development beginning from a consideration of traffic control methods at an isolated link intersection to a discussion of very complex approaches to dynamic traffic signal

control and optimization in a complex city network. Researchers that developed standard methods for control of signalized intersection reviewed by Gartner and Stamatiadis (2009) could not know the nucleation nature of traffic breakdown at the signal predicted only in 2011 and reviewed in Kerner (2017c). However, Gartner and Stamatiadis in their later developments of the standard models for dynamic traffic signal control (Zhang et al. 2015) have also not discussed the nucleation nature of traffic breakdown at the signal.

It should be emphasized that the nucleation nature of traffic breakdown has been empirically proven in studies of real field traffic data measured at highway bottlenecks only. In contrast with traffic flow measurements available for highway bottlenecks, the quality of measured data at signalized intersection is currently still not satisfactory enough to study the nucleation features of transition from under- to over-saturated traffic (traffic breakdown) at the signal. This will be an interesting task for a study of single vehicle data in the future.

A critical analysis of standard traffic flow models related to the state-of-the-art in traffic and transportation research has been made in the article by Kerner (2018a). In this article, it will be shown that the standard traffic flow models are inconsistent with the empirical nucleation nature of traffic breakdown at network bottlenecks. A probabilistic theory of traffic breakdown is considered in the review article of Kerner and Klenov (2018).

Cellular automaton (CA) traffic flow models are a class of microscopic traffic flow models. In the CA models, the time and space are discrete. The CA models have the advantage of flexible evolution rules of vehicular motion and high computation efficiency. The development of three-phase CA traffic flow models is an important contribution to the three-phase traffic theory. A review of cellular automaton traffic flow models in the framework of the three-phase theory has been made by Tian et al. (2018).

It is assumed that future vehicular traffic is mixed traffic flow consisting of autonomous driving vehicles that are randomly distributed

between human driving vehicles. An autonomous driving vehicle is also called a self-driving vehicle or an automated driving vehicle or else an automatic driving vehicle. One of the problems for autonomous driving is as follows. An autonomous driving vehicle that moves unexpectedly for human drivers can be considered as an “obstacle” for the drivers (e.g., such a case can occur by vehicle merging at bottlenecks, by lane changing, by overtaking, and by tuning). This can decrease traffic safety. This can also lead to large local disturbances in traffic flow. In turn, the disturbances can initiate traffic breakdown. Thus, autonomous driving systems should be developed in the future that learn from behavior of human drivers. Autonomous driving strategy that learns from human behavior incorporated in the three-phase traffic theory is reviewed in Kerner (2018c). In this entry, we discuss possible advantages of the autonomous driving strategy in the framework of three-phase traffic theory versus classical (standard) autonomous driving strategy.

Traffic congestion resulting from traffic breakdown in a traffic network exhibits a complex spatiotemporal behavior that was studied by several generations of scientists. Empirical spatiotemporal features of traffic congested patterns resulting from traffic breakdown are considered in the review article by Kerner (2018b). Empirical observations of traffic congestion discussed in this article show that phenomena of traffic breakdown and the subsequent evolution of resulting congested patterns are associated with diverse phase transitions and complex spatiotemporal self-organization effects in vehicular traffic.

Engineering applications of the three-phase theory for the reconstruction of spatiotemporal traffic patterns and traffic prediction are discussed by Rehborn et al. (2017). In the dynamic traffic management models, a reliable prognosis of traffic is also required. As shown in the article (Rehborn et al. 2017), there are two main approaches to traffic prognosis: (i) “physics-of-traffic,” i.e., traffic prognosis models and methods based on the understanding of the nature of empirical spatiotemporal features of traffic patterns; and (ii) methods of mathematical statistics or artificial intelligence, which are also called “data mining”

methods, in which the understanding of traffic patterns is not necessarily needed. Data mining models for traffic prognosis do not usually require the understanding of traffic data, which these methods use for traffic prediction. Using a huge number of historical traffic data, such a method learns predictable features of traffic data without the pretension of their understanding. The data mining methods used for traffic prognosis include statistical methods based on regression, a wavelet approach, or filtering models as well as neural networks. It should be noted that there are many articles in this Encyclopedia, in which these and other methods of artificial intelligence are reviewed. The physics-of-traffic approach to traffic prognosis applied in Rehborn et al. (2017) is based on the understanding of measurements of traffic variables made in space and time as well as on simulations of the traffic variables associated with empirical data. The key element for a successful prediction concept lies in the correct and reliable understanding of the empirical features of the traffic process and their reproducible characteristics. Furthermore, this understanding of real traffic should be incorporated in a traffic flow model, which should explain and predict the empirical traffic features. Therefore, in the physics-of-traffic approach, a traffic flow prediction model is based on empirical features of phase transitions and/or resulting congested patterns; current and historical measured traffic data are used in the model, which reconstructs traffic phases and makes the tracking and prediction of the propagation of traffic congestion in a traffic network.

The understanding of driver behavior in traffic flow is important for the development of microscopic traffic flow models. In a review article of Lubashevsky and Morimura (2018), the authors study how drivers perceive the current situation in the vicinity of their cars, their ability to anticipate the car arrangement in the near future, the individual goals the drivers intend to achieve in future, the gained experience of driving, fatigue, and the social norms of driving. Leaving aside the philosophical long-term debates about the reducibility of the mental to the physical, Lubashevsky and Morimura confine their consideration to the

approach dealing with physical states of car motion and the driver perception of these states as they appear in the mind. The approach of Lubashevsky and Morimura may be categorized as phenomenology of human actions in controlling unstable mechanical systems and belongs to the scope of novel interdisciplinary branch of science called physics of mind (see references to the field *physics of mind* in Lubashevsky and Morimura (2018)). In article by Lubashevsky and Morimura (2018), the authors employ the gist of phenomenological approach for constructing a model for the car-following, a paradigmatic example of intentional human actions in governing mechanical systems.

For an efficient transportation, it is important to operate buses and trams on time. The bus schedule is closely related to the dynamic motion of buses. In a review of Nagatani (2017), the nonlinear maps for describing the dynamics of shuttle buses in the transportation system have been introduced and described. The complex motion of the buses is explained by nonlinear-map models. The transportation system of shuttle buses without passing is similar to that of the trams. The transport of elevators is also similar to that of shuttle buses with freely passing. The complex dynamics of a single bus is described in terms of the piecewise map, the delayed map, the extended circle map, and the combined map. The dynamics of a few buses is described by the model of freely passing buses, the model of no passing buses, and the model of increase or decrease of buses. The nonlinear-map models are useful to make an accurate estimate of the arrival time in the bus transportation.

Modeling approaches to vehicular traffic management discussed in the above mentioned review articles of the Encyclopedia are also very important for studies of air traffic and pedestrian traffic.

Currently, we can observe a continuous growth of publications devoted to modeling approaches to pedestrian traffic and crowd dynamics. In particular, the interest to pedestrian traffic is associated with necessity in the development of effective evacuation strategies required for the relocation of people to safely escape hazardous disaster impacts. During the evacuation, destinations are chosen to escape the hazardous disaster

impacts as quickly as possible. This is a peculiarity of the evacuation process in comparison with vehicle traffic in which every vehicle has usually a defined destination.

Achievements and problems in the understanding of pedestrian and evacuation dynamics are reviewed in the articles by Chraïbi et al. (2018) as well as by Boltes et al. (2018). In these review articles, various aspects of the modeling of pedestrian interactions during evacuation as well as a comparison of the associated theoretical and empirical results are considered.

Based on an example of Australian fire zone research, Goudie (2017) considers evacuation as a communication and social phenomenon. Goudie's article describes the influence of communication and social aspects on possible evacuation scenarios.

Air traffic continues to grow at a steady pace in the world. For this reason, air traffic is undergoing major changes both in developed and developing countries. The demand for air traffic in the world is expected to grow to several times the current levels of traffic in the next few decades. Recently, methods have been developed to study the network characteristics of large complex systems in many natural science and engineering fields. Several networks exhibit a scale-free property in the sense that the probabilistic distribution of their nodes as a function of connections decreases slower than an exponential. These networks are characterized by the fact that a small number of components have a disproportionate influence on the performance of the network. Therefore, the discussion of air traffic dynamics as well as approaches to air traffic modeling made in the review article of Sridhar and Sheth (2019) is an important contribution to the fields of traffic management.

Acknowledgments I would like to thank Michael Schreckenberg, Hubert Rehborn, Craig Davis, Ihor Lubashevsky, Sergey Klenov, Micha Koller, Sven-Eric Molzahn, Dominik Wegerle, and Yildirim Dülger for many useful comments. I thank our partners for their support in the projects “UR:BAN – Urban Space: User oriented assistance systems and network management” and “MEC-View – Object detection for automated driving based on Mobile Edge Computing,” funded by the German Federal Ministry of Economic Affairs and Energy.

Bibliography

- Ashton WD (1966) The theory of traffic flow. Wiley/Methuen & Co, London/New York
- Barceló J (ed) (2010) Fundamentals of traffic simulation. International series in operations research and management science, vol 145. Springer, Berlin
- Bell MGH, Iida Y (1997) Transportation network analysis. Wiley, Hoboken Incorporated
- Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. *Math Models Methods Appl Sci* 12:1801–1843
- Boltes M, Zhang J, Tordeux A, Schadschneider A, Seyfried A (2018) Pedestrian and evacuation dynamics: empirical results. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Brackstone M, McDonald M (1999) Car-following: a historical review. *Transp Res F* 2:181–196
- Brockfeld E, Kühne RD, Skabardonis A, Wagner P (2003) Toward benchmarking of microscopic traffic flow models. *Trans Res Rec* 1852:124–129
- Buslaev AP, Tatashev AG, Yashina MV (2013) Mathematical physics of traffic. Technical Center Publisher, Moscow (in Russian)
- Castiglione J, Bradley M, Gliebe J (2015) Activity-based travel demand models: a primer. SHRP 2 Report S2-C46-RR-1. Transportation Research Board, Washington, DC
- Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6:165–184
- Chiu Y-C, Bottom J, Mahut M, Paz A, Balakrishna R, Waller T, Hicks J (2011) Dynamic traffic assignment. A primer. *Trans Res Circular E-C153* <http://onlinepubs.trb.org/onlinepubs/circulars/ec153.pdf>
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
- Chraïbi M, Tordeux A, Schadschneider A, Seyfried A (2018) Pedestrian and evacuation dynamics – modelling. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Daganzo CF (1997) Fundamentals of transportation and traffic operations. Elsevier Science, New York
- Drew DR (1968) Traffic flow theory and control. McGraw-Hill Book, New York
- Edie LC (1954) Traffic delays at toll booths. *Oper Res* 2:107–138
- Elefteriadou L (2014) An introduction to traffic flow theory. Springer optimization and its applications, vol 84. Springer, Berlin
- Ferrara A, Saccone S, Siri S (2018) Freeway traffic modelling and control. Springer, Berlin
- Friesz TL, Bernstein D (2016) Foundations of network optimization and games, Complex networks and dynamic systems, vol 3. Springer, New York/Berlin
- Gartner NH, Stamatiadis C (2009) Traffic networks, optimization and control of urban. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9470–9500
- Gartner NH, Messer CJ, Rathi A (eds) (1997) Special report 165: revised monograph on traffic flow theory. Transportation Research Board, Washington, DC
- Gartner NH, Messer CJ, Rathi A (eds) (2001) Traffic flow theory: a state-of-the-art report. Transportation Research Board, Washington, DC
- Gasnikov AV, Klenov SL, Nurminski EA, Kholodov YA, Shamray NB (2013) Introduction to mathematical simulations of traffic flow. MCNMO, Moscow (in Russian)
- Gazis DC (2002) Traffic theory. Springer, Berlin
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7:499–505
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
- Gerlough DL, Huber MJ (1975) Traffic flow theory. Special Report 165. Transportation Research Board, National Research Council. Washington, DC
- Goudie D (2017) Natural disasters and evacuations as a communication and social phenomenon. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Goulias K (2009) Travel behavior and demand analysis and prediction. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9536–9565
- Greenshields BD, Bibbins JR, Channing WS, Miller HH (1935) A study of traffic capacity. *Highw Res Board Proc* 14:448–477
- Haight FA (1963) Mathematical theories of traffic flow. Academic, New York
- Hegyi A, Bellemans T, De Schutter B (2017) Freeway traffic management and control. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Helbing D (2001) Traffic and related self-driven many particle systems. *Rev Mod Phys* 73:1067–1141
- Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7:86–106
- Highway Capacity Manual (2000) National research council. Transportation Research Board, Washington, DC
- Highway Capacity Manual (2016) National research council. Transportation Research Board, Washington, DC
- Horn A, Nagel K, Axhausen KW (eds) (2016) The multi-agent transport simulation MATSim. Ubiquity, London. <https://doi.org/10.5334/baw>. URL: <http://matsim.org/the-book>
- Kachroo P, Özbay KMA (2018) Feedback control theory for dynamic traffic assignment. Springer, Berlin
- Kerner BS (2004) The physics of traffic. Springer, Berlin/New York
- Kerner BS (2009) Introduction to modern traffic flow theory and control. Springer, Berlin/New York
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Physica A* 392:5261–5282

- Kerner BS (2015) Failure of classical traffic flow theories: a critical review. *Elektrotech Inf* 132:417–433
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Physica A* 450:700–747
- Kerner BS (2017a) Breakdown minimization principle versus Wardrops equilibria for dynamic traffic assignment and control in traffic and transportation networks: a critical mini-review. *Physica A* 466:626–662
- Kerner BS (2017b) Breakdown in traffic networks. Springer, Berlin/New York
- Kerner BS (2017c) Traffic networks, breakdown in. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin. <https://doi.org/10.1007/978-3-642-27737-5701-1>
- Kerner BS (2018a) Traffic breakdown, modeling approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin. <https://doi.org/10.1007/978-3-642-27737-5559-2>
- Kerner BS (2018b) Traffic congestion, spatiotemporal features of. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin. <https://doi.org/10.1007/978-3-642-27737-5560-2>
- Kerner BS (2018c) Autonomous driving in the framework of three-phase traffic theory. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin. https://doi.org/10.1007/978-3-642-27737-5_724-1
- Kerner BS, Klenov SL (2018) Traffic breakdown, mathematical probabilistic approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin. https://doi.org/10.1007/978-3-642-27737-5_558-3
- Kessels F (2019) Traffic flow modelling. Springer, Berlin
- Kometani E, Sasaki T (1958) On the stability of traffic flow (report-I). *Oper Res Soc Jpn* 2:11–26
- Kometani E, Sasaki T (1959) A safety index for traffic with linear spacing. *Oper Res* 7:704–720
- Kometani E, Sasaki T (1961a) Car following theory and stability limit of traffic volume. *Oper Res Soc Jpn* 3:176–190
- Kometani E, Sasaki T (1961b) Dynamic behavior of traffic with a nonlinear spacing-speed relationship. In: *Proceedings of symposium on the theory of traffic flow*. Elsevier, Amsterdam, pp 105–119
- Kuhn TS (2012) *The structure of scientific revolutions*, 4th edn. The University of Chicago Press, Chicago/London
- Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
- Lighthill MJ, Whitham GB (1995) On kinematic waves. I flow movement in long rivers. II a theory of traffic flow on long crowded roads. *Proc Roy Soc A* 229:281–345
- Lubashevsky I, Morimura K (2018) Physics of mind and car-following problem. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419:1–64
- Mahmassani HS (2001) *Dynamic network traffic assignment and simulation methodology for advanced system management applications*. *Netw Spat Econ* 1:267–292
- Mahnke R, Kaupužs J, Lubashevsky I (2009) *Physics of stochastic processes: how randomness acts in time*. Wiley-VCH, Weinheim
- Mannering FL, Kilareski WP (1998) *Principles of highway engineering and traffic analysis*, 2nd edn. Wiley, New York
- May AD (1990) *Traffic flow fundamentals*. Prentice-Hall, Englewood Cliffs
- Moskowitz K (1954) Waiting for a gap in a traffic stream. *Proc Highw Res Board* 33:385–395
- Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
- Nagatani T (2017) Complex dynamics of bus, tram, and elevator delays in transportation systems. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Nagel K, Wagner P, Woessler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–716
- Newell GF (1961) Nonlinear effects in the dynamics of car following. *Oper Res* 9:209–229
- Newell GF (1963) Instability in dense highway traffic, a review. In: *Proceedings of the second international symposium on traffic road traffic flow*. OECD, London, pp 73–83
- Newell GF (1982) *Applications of queuing theory*. Chapman Hall, London
- Papageorgiou M, Papamichail I (2008) Overview of traffic signal operation policies for ramp metering. *Transp Res Rec* 2047:2836
- Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. *Netw Spat Econ* 1:233–265
- Piccoli B, Tosin A (2009) Vehicular traffic: a review of continuum mathematical models. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9727–9749
- Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24:274–281
- Prigogine I (1961) A Boltzmann-like approach to the statistical theory of traffic flow. *Proceedings of the first international symposium on the theory of traffic flow* (Warren, Mich., 1959). Elsevier, Amsterdam/New York, pp 158–164
- Prigogine I, Herman R (1971) *Kinetic theory of vehicular traffic*. American Elsevier, New York
- Rakha H, Tawfik A (2009) Traffic networks: dynamic traffic routing, assignment, and assessment. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9429–9470
- Ran B, Boyce D (1996) *Modeling dynamic transportation networks*. Springer, Berlin
- Rehborn H, Klenov SL, Koller M (2017) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin

- Rempe F, Franeck P, Fastenrath U, Bogenberger K (2016) Online freeway traffic estimation with real floating car data. In: Proceedings of 2016 I.E. 19th international conference on ITS, pp 1838–1843
- Rempe F, Franeck P, Fastenrath U, Bogenberger K (2017) A phase-based smoothing method for accurate traffic speed estimation with floating car data. *Trans Res C* 85:644–663
- Reuschel A (1950a) Vehicle movements in a platoon. *Österreichisches Ing-Arch* 4:193–215 (In German)
- Reuschel A (1950b) Vehicle movements in a platoon with uniform acceleration or deceleration of the lead vehicle. *Z. Österreichischen Ing.-Archit.-Ver* 95: 59–62, 73–77 (1950) (In German)
- Richards PI (1956) Shockwaves on the highway. *Oper Res* 4:42–51
- Roess RP, Prassas ES (2014) The highway capacity manual: a conceptual and research history. Springer, Berlin
- Saifuzzaman M, Zheng Z (2014) Incorporating human factors in car-following models: a review of recent developments and research needs. *Transp Res C* 48:379–403
- Schadschneider A, Chowdhury D, Nishinari K (2011) Stochastic transport in complex systems. Elsevier Science, New York
- Seo T, Bayen AM, Kusakabe T, Asakura Y (2017) Traffic state estimation on highway: a comprehensive survey. *Annu Rev Control* 43:128–151
- Sheffi Y (1984) Urban transportation networks: equilibrium analysis with mathematical programming methods. Prentice-Hall, Englewood Cliffs
- Shvetsov VI (2003) Mathematical modeling of traffic flow. *Autom Remote Control* 64:1651–1689
- Sridhar B, Sheth K (2019) Air traffic control, complex dynamics of. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer, Berlin
- Tian J-F, Zhu C, Jiang R (2018) Cellular automaton models in the framework of three-phase traffic theory. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer, Berlin
- TRANSIMS (2017) U.S. Department of Transportation, Federal Highway Administration, Washington, DC. <https://www.fhwa.dot.gov/planning/tmip/resources/transims/>
- TRANSIMS Open Source (2013) Getting started with TRANSIMS, webpage, <http://code.google.com/p/transims/wiki/GettingStarted>
- Treiber M, Kesting A (2013) Traffic flow dynamics. Springer, Berlin
- Wardrop JG (1952) Some theoretical aspects of road traffic research. In: Proceedings of the institute of civil engineers part II, vol 1, pp 325–378
- Webster FV (1958) Traffic signal settings. HMSO, London
- Zhang C, Xie Y-C, Gartner NH, Stamatiadis C, Arsava T (2015) AM-band: an asymmetrical multi-band model for arterial traffic signal coordination. *Transp Res C* 58:515–531

Breakdown in Traffic Networks

Boris S. Kerner
Physics of Transport and Traffic, University
Duisburg-Essen, Duisburg, Germany

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Empirical Nucleation Nature of Traffic
Breakdown: Empirical Fundamental of
Transportation Science
Failure of Classical Traffic Flow Models and
Theories
Theoretical Fundamental of Transportation
Science: The Three-Phase Theory
Incommensurability of Classical Traffic Theories
with Three-Phase Theory
Classical Understanding of Stochastic Highway
Capacity
Failure of Intelligent Transportation Systems
(ITS) Based on Classical Traffic Theories
Congested Pattern Control Approach Based on
Three-Phase Theory
Decrease in Probability of Traffic Breakdown in
Networks Through Automated Driving
Vehicles
Deterioration of Performance of Traffic System
Through Automated Driving Vehicles
Automated Driving Based on Three-Phase
Theory
Random Time-Delayed Breakdown in City
Traffic
Theoretical Fundamental of Transportation
Science: Breakdown Minimization
(BM) Principle
Network Throughput Maximization Approach
A Physical Measure of Traffic and Transportation
Networks: Network Capacity

Deterioration of Traffic in Networks Through
Standard Dynamic Traffic Assignment
Conclusions: Future Dynamic Traffic Assignment
and Control in Networks
Bibliography

Glossary

Bottleneck Traffic breakdown leading to the onset of congestion in a traffic network occurs usually at a network bottleneck. Road bottlenecks are caused, for example, by on- and off-ramps, road gradients, roadworks, a decrease in the number of road lines (in the flow direction), traffic signal, etc.

Breakdown Minimization (BM) Principle The BM principle states that the optimum of a traffic or transportation network with N bottlenecks is reached when dynamic traffic assignment, optimization, and control are performed in the network in such a way that the probability P_{net} for the occurrence of traffic breakdown in at least one of the network bottlenecks during a given time interval T_{ob} for observing traffic flow reaches the minimum possible value. The BM principle is equivalent to the maximization of the probability that during the time interval T_{ob} , traffic breakdown occurs at none of the network bottlenecks.

Congested Traffic Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is still possible in free flow.

Congested Pattern Control Approach In congested pattern control approach, *no control* of traffic flow at a network bottleneck is realized as long as free flow is realized at the bottleneck. This means that the occurrence of random traffic breakdown is permitted to occur at the bottleneck: Congestion at the bottleneck is allowed to set in. The main idea of this approach is as follows. Only after traffic breakdown has occurred, traffic control starts. Due to the congested pattern control approach,

either free flow recovers at the bottleneck or traffic congestion is localized at the bottleneck.

Constrain “Alternative Network Routes (Paths)” By any application of the breakdown minimization (BM) principle, there is a constrain: “alternative network routes (paths).” The constrain “alternative routes” prevents the use of very long routes in dynamic traffic assignment with the use of the BM principle.

Dynamic Traffic Assignment Dynamic traffic assignment is assignment in a traffic or transportation network considering the temporal dimension of the assignment problem.

Empirical Fundamental of Transportation Science The nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck is an empirical fundamental of transportation science.

Empirical Traffic Breakdown Empirical traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck can occur in a metastable state of free flow with respect to the $F \rightarrow S$ transition. In the metastable state of free flow, traffic breakdown occurs only if a nucleus required for the breakdown appears. Empirical traffic breakdown ($F \rightarrow S$ transition) at the bottleneck can be either spontaneous or induced.

Free Flow Free traffic flow (free flow for short) is usually observed, when the vehicle density in traffic is small enough. The flow rate increases in free flow with the increase in the vehicle density, whereas the average vehicle speed is a decreasing density function. In accordance with real field traffic data, there is a limit in the increase in the vehicle density in free flow: In the associated limit state of free flow, the flow rate and density reach their maximum values, while the average speed reaches a minimum value that is still possible in free flow.

$F \rightarrow S$ Transition In all known observations, traffic breakdown at a highway bottleneck is a phase transition from the free flow phase to synchronized flow phase called an $F \rightarrow S$ transition. The terms *traffic breakdown* and *$F \rightarrow S$ transition* are synonyms related to the same phenomenon of the onset of congestion in free flow.

Incommensurability of Classical Traffic Theories with the Three-Phase Traffic Theory The

three-phase theory has been introduced for the explanation of the empirical free flow metastability with respect to an $F \rightarrow S$ transition at a highway bottleneck. This metastability of free flow has no sense for classical traffic and transportation theories. In accordance with the classical theory by Kuhn about scientific revolutions, this shows the incommensurability of the three-phase theory and the classical traffic flow theories. The term “incommensurable” has been introduced by Kuhn to explain a paradigm shift in the scientific field.

Induced Traffic Breakdown (Induced $F \rightarrow S$ Transition) at Bottleneck An induced traffic breakdown (induced $F \rightarrow S$ transition) in metastable free flow at a bottleneck is traffic breakdown that is induced at the bottleneck by the propagation of a moving spatiotemporal congested traffic pattern. This congested pattern has occurred *earlier* than the time instant of traffic breakdown at the bottleneck and at a *different* road location (e.g., at another bottleneck) than the bottleneck location. When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck. The empirical induced $F \rightarrow S$ transition at the bottleneck proves the empirical nucleation nature of traffic breakdown at the bottleneck. The empirical nucleation nature of traffic breakdown at a highway bottleneck can be considered the empirical fundamental of transportation science.

Maximum Highway Capacity A maximum highway capacity is equal to the flow rate at a highway bottleneck that separates metastable states of free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck from unstable states of free flow: When the flow rate is equal to or larger than the maximum highway capacity, free flow is unstable with respect to traffic breakdown at the bottleneck. The maximum highway capacity is a stochastic value.

Metastable Free Flow with Respect to Traffic Breakdown ($F \rightarrow S$ Transition) Metastable free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck is free flow in which small enough local speed

(and density) disturbances decay over time, i.e., the small local disturbances cause no traffic breakdown. However, if a large enough local disturbance appears in metastable free flow at the bottleneck, the $F \rightarrow S$ transition does occur. In the latter case, the local disturbance is called a nucleus for the $F \rightarrow S$ transition.

Minimum Highway Capacity A minimum highway capacity is equal to the flow rate at a highway bottleneck that separates stable states of free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck from metastable states of free flow at the bottleneck. When the flow rate is smaller than the minimum highway capacity, free flow is stable with respect to traffic breakdown at the bottleneck. No traffic breakdown can occur at the bottleneck. This is independent of the amplitude of a time-limited local disturbance in free flow at the bottleneck. The minimum highway capacity is a stochastic value.

Moving Traffic Jam A moving traffic jam (moving jam for short) is a localized congested traffic pattern that moves upstream in traffic flow. Within the moving jam the average vehicle speed is very low (sometimes as low as zero), and the density is very high. The moving jam is spatially restricted by the downstream jam front and upstream jam front. Within the downstream jam front vehicles accelerate from low speed states within the jam to higher speeds in traffic flow downstream of the moving jam. Within the upstream jam front vehicles must slow down to the speed within the jam.

Network Throughput Maximization Approach A network throughput maximization approach is an approach to dynamic traffic assignment and control in a traffic or transportation network. The network throughput maximization approach is an application of the BM principle that maximizes the network throughput while keeping free flow conditions in the whole network. The network throughput maximization approach maintains the condition $P_{\text{net}} = 0$, where P_{net} is the probability of traffic breakdown in the network. Therefore, the network throughput maximization approach is also called the BM principle for “zero breakdown probability.”

Through the maximization of the network throughput with the network throughput maximization approach, no traffic breakdown can occur in the traffic network.

Network Capacity A network capacity denoted by C_{net} is a measure (or “metric”) of a traffic or transportation network. Under a steady state analysis of the network, as long as the total network inflow rate is smaller than the network capacity, traffic breakdown cannot occur in the network. The network capacity is a stochastic value.

Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition) The empirical nucleation nature of the $F \rightarrow S$ transition is as follows. Traffic breakdown at a highway bottleneck is a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition). The $F \rightarrow S$ transition occurs in a metastable state of free flow with respect to this transition at a highway bottleneck. There can be many different local speed disturbances of the speed decrease in metastable free flow at the bottleneck. Small enough local speed disturbances in metastable free flow at the bottleneck decay. However, large enough local speed disturbances in metastable free flow lead to the $F \rightarrow S$ transition at the bottleneck. A local speed disturbance occurring in the metastable free flow that leads to the $F \rightarrow S$ transition is called a nucleus for the $F \rightarrow S$ transition. The nucleation nature of the $F \rightarrow S$ transition is explained by the nucleation feature of the $S \rightarrow F$ instability: The $S \rightarrow F$ instability governs traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck.

Nucleus for Traffic Breakdown A nucleus for traffic breakdown ($F \rightarrow S$ transition) in metastable free flow is a local speed (density) disturbance in free flow whose occurrence causes traffic breakdown at a highway bottleneck. The average speed within the nucleus for traffic breakdown is lower than the average speed in free flow outside the nucleus. There can be local speed (density) disturbances with different disturbance amplitudes that can be nuclei for traffic breakdown in metastable free flow at the bottleneck. A local speed (density) disturbance with the minimum possible disturbance

amplitude (critical disturbance) whose occurrence causes traffic breakdown in metastable free flow at the bottleneck is called a critical nucleus for traffic breakdown. Thus, a critical disturbance with respect to the $F \rightarrow S$ transition in free flow is a synonym of a critical nucleus for traffic breakdown. The amplitude of local speed (density) disturbance that is the critical nucleus for traffic breakdown depends on the flow rate in metastable free flow at the bottleneck.

Probability of Traffic Breakdown in Traffic or Transportation Network The probability of traffic breakdown in a traffic or transportation network P_{net} is the probability that traffic breakdown occurs during a given time interval for observing traffic flow at least at one of the network bottlenecks (probability of traffic breakdown in the network, for short).

$S \rightarrow F$ Instability An $S \rightarrow F$ instability is an instability of synchronized flow. Due to the $S \rightarrow F$ instability, an initial speed disturbance of a local increase in the speed within synchronized flow (S) at the bottleneck transforms into a growing speed wave of the increase in the speed (growing acceleration wave) that propagates upstream within synchronized flow and leads to free flow (F) at the bottleneck: The $S \rightarrow F$ instability results in an $S \rightarrow F$ transition. The $S \rightarrow F$ instability is caused by driver over-acceleration. The $S \rightarrow F$ instability occurs due to a finite time delay in driver over-acceleration.

Saturation Flow Rate The maximum flow rate in the outflow from a vehicle queue at a traffic signal during the green signal phase determines the saturation flow rate when the following conditions are satisfied: (i) The vehicles are in a standstill within the queue, (ii) vehicles escape freely from the queue at its downstream front, and (iii) the vehicles can accelerate to the maximum vehicle speed permitted in city traffic downstream of the signal. In empirical observations, the saturation flow rate is a characteristic flow rate that does not depend on distances between vehicles that are in a standstill in the queue, on the queue length, or on the arrival flow rate upstream of the queue.

However, the saturation flow rate can depend considerably on traffic parameters (percentage of long vehicles (trucks), weather, road conditions, etc.).

Standard Dynamic Traffic Assignment Standard dynamic traffic assignment is related to approaches to dynamic traffic assignment with the use of the Wardrop's UE or SO, whose objective is the minimization of route travel times and/or other travel costs in a traffic and transportation network. The term "standard" emphasizes that the approaches to dynamic traffic assignment are related to the state of the art in traffic and transportation research.

Synchronized Flow Traffic Phase In the three-phase theory, the following definition (criterion) [S] of the synchronized flow phase (synchronized flow for short) in congested traffic is made. In contrast to the wide moving jam phase, the downstream front of the synchronized flow phase does *not* maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often fixed at a highway bottleneck. In other words, synchronized flow does not exhibit the characteristic jam feature [J].

Theoretical Fundamentals of Transportation Science The three-phase theory is the theoretical fundamental of transportation science that explains the empirical nucleation nature of traffic breakdown. The BM principle is the theoretical fundamental of transportation science that permits to maximize the network throughput preventing traffic breakdown in the network.

Three-Phase Traffic Theory In the three-phase traffic theory (three-phase theory for short) introduced by the author, besides the free flow phase, there are two phases in congested traffic, the synchronized flow (S) and wide moving jam (J) phases. The synchronized flow and wide moving jam phases in congested traffic are defined through the use of empirical spatiotemporal criteria (definitions) [S] and [J], respectively. The three-phase theory is a qualitative theory that explains empirical traffic breakdown and resulting empirical congested

traffic patterns based on the criteria [S] and [J] for traffic phases as well as on hypotheses of this theory. The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown at highway bottlenecks. In the three-phase theory, the empirical nucleation nature of traffic breakdown at a highway bottleneck is explained by the metastability of free flow with respect to an $F \rightarrow S$ transition at the bottleneck.

Time Delay of Traffic Breakdown at Highway Bottleneck There is a time delay of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck. The time delay is a random value. A random time delay of traffic breakdown at the bottleneck is caused by a sequence of $F \rightarrow S \rightarrow F$ transitions occurring at the bottleneck: The $S \rightarrow F$ transition in this sequence prevents traffic breakdown at the bottleneck. The $S \rightarrow F$ transition is determined by the $S \rightarrow F$ instability at the bottleneck. Because the $S \rightarrow F$ instability exhibits the nucleation nature, the $S \rightarrow F$ instability does not occur, if no nucleus required for the $S \rightarrow F$ instability appears in synchronized flow at the bottleneck. In this case, traffic breakdown does occur at the bottleneck. The occurrence of the nucleus required for the $S \rightarrow F$ instability in synchronized flow at the bottleneck is a random event. This explains that there is a random time delay of traffic breakdown at the bottleneck. This explains also a result of the three-phase theory that the $S \rightarrow F$ instability governs traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. For this reason, traffic breakdown at the bottleneck can be called a time-delayed traffic breakdown at the bottleneck.

Time Delay of Traffic Breakdown at Traffic Signal Traffic breakdown at the signal is a transition from undersaturated traffic to oversaturated traffic at the signal. There is a time delay of traffic breakdown at traffic signal. The time delay is a random value. The time delay of traffic breakdown at traffic signal is associated with the metastability of undersaturated traffic at the signal with respect to the transition to oversaturated traffic at the signal. For these

reasons, traffic breakdown at the signal in city traffic can be called a time-delayed transition from undersaturated traffic to oversaturated traffic at the signal.

Traffic Assignment Traffic assignment is a procedure used to find the network link flow rates from the original-destination demand.

Traffic Breakdown at Highway Bottleneck

Traffic breakdown at a highway bottleneck is the phenomenon of the onset of congestion in an initial free flow at the bottleneck. In all observations, traffic breakdown is an $F \rightarrow S$ transition at the bottleneck. The terms *$F \rightarrow S$ transition*, *breakdown phenomenon*, *traffic breakdown*, and *speed breakdown* are synonyms related to the same effect: the onset of congestion in free flow at the bottleneck.

Traffic Flow Model A traffic flow model is devoted to the explanation and simulation of traffic flow phenomena, which are observed in measured data of real traffic flow, and to the prediction of new traffic flow phenomena that could be found in real traffic flow.

Wardrop's Equilibria The Wardrop's user equilibrium (UE) and system optimum (SO) are also called the Wardrop's equilibria or the Wardrop's principles.

Wardrop's System Optimum (SO) The Wardrop's SO states that the average travel times of all vehicles in a network are minimum, which implies that aggregate vehicle-hours spent in travel are also minimum. In other words, in accordance with Wardrop's SO, the network-wide travel time should be a minimum.

Wardrop's User Equilibrium (UE) The Wardrop's UE states that traffic on a network distributes itself in such a way that the travel times on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel times.

Wide Moving Jam Traffic Phase In the three-phase theory, the following definition (criterion) [J] of the wide moving jam traffic phase (wide moving jam for short) in congested traffic is made. A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the

jam propagates through any other traffic states or highway bottlenecks. This is the characteristic feature [J] of the wide moving jam phase.

Z-Characteristic for Traffic Breakdown A Z-characteristic of traffic breakdown is a macroscopic characteristic of traffic breakdown at a highway bottleneck. The Z-characteristic is built by metastable states of the traffic phases F and S together with a branch for the average speed within critical nuclei for traffic breakdown ($F \rightarrow S$ transition) at the bottleneck.

Definition of the Subject and Its Importance

A study of spatiotemporal vehicular traffic phenomena in traffic networks is necessary for the organization of high-quality mobility in any country of the world.

There can be either free traffic flow (free flow (F) for short) or congested traffic in a traffic network. In free flow, a vehicle can move almost freely during long time intervals because interactions with other vehicles are not very strong. Contrarily, in congested traffic there are very strong interactions between vehicles. For this reason, the average vehicle speed is considerably smaller in congested traffic than the average speed in free flow. This results in a considerable increase in travel times and fuel consumption in congested traffic in comparison with the associated travel costs in free flow in the network. This explains why the main aim of traffic network management, traffic control, dynamic traffic assignment, and other traffic engineering applications through the use of intelligent transportation systems (ITS) in a traffic network is to keep free flow conditions in the network.

In real traffic, congested traffic occurs usually due to traffic breakdown at network bottlenecks. In other words, traffic breakdown is a transition from free flow to congested traffic at a network bottleneck. The bottleneck can be a result of road works, on- and off-ramps, a decrease in the number of road lanes, road curves and road gradients, traffic signal, etc. The probability of traffic breakdown at the bottleneck is considerably larger than that outside the bottleneck.

Free flow conditions in the network can be satisfied only when traffic breakdown in a traffic network can be prevented. This explains the fundamental importance of the understanding of the nature of traffic breakdown at network bottlenecks for the development of any reliable approaches and methods of network management, cooperative driving, automated driving, traffic control, dynamic traffic assignment, and other ITS-applications in a traffic network.

Introduction

The most important task of empirical studies of vehicular traffic in a traffic network is to find and understand empirical features of traffic breakdown in the network (see, e.g., Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Ceder 1999; Daganzo 1997; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Greenshields et al. 1935; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; Lesort 1996; Lorenz and Elefteriadou 2000; Mahmassani 2005; May 1990; Persaud et al. 1998; Taylor 2002; Zurlinden 2003). This is because congested traffic resulting from traffic breakdown causes a considerable increase in travel time, fuel consumption, CO₂ emission, and other travel costs. Therefore, any traffic and transportation theory applied for the development of automated driving vehicles, reliable methods of dynamic traffic assignment and control, and any other ITS-applications (ITS – intelligent transportation systems) should be consistent with the empirical features of traffic breakdown at a network bottleneck. Road bottlenecks are caused for example by on- and off-ramps, road gradients, roadworks, a decrease in the number of road lines (in the flow direction), traffic signal, etc. (see, e.g., Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Ceder 1999; Daganzo 1997; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Greenshields et al. 1935; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; Lesort 1996; Lorenz and Elefteriadou 2000; Mahmassani 2005; May 1990; Persaud et al. 1998; Taylor 2002; Zurlinden 2003).

In real traffic data, it has been found that traffic breakdown at a highway bottleneck is a phase transition from free flow (F) to congested traffic whose downstream front is usually fixed at the bottleneck (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990).

In the three-phase theory, such congested traffic has been called synchronized flow (S). This means that traffic breakdown at the highway bottleneck is an $F \rightarrow S$ transition. It has turned out that in real traffic data, traffic breakdown occurs in a *metastable state* of free flow with respect to the $F \rightarrow S$ transition at the bottleneck (Kerner 1998a, b, 1999a, b, c).

The metastability of free flow with respect to the $F \rightarrow S$ transition is as follows. There can be many local speed (density, flow rate) disturbances in free flow at a bottleneck. Amplitudes of the local disturbances can be very different. When a local disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown ($F \rightarrow S$ transition) occurs. Such a local disturbance resulting in the breakdown is called *nucleus* for the breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays; as a result, no traffic breakdown occurs. In other words, empirical traffic breakdown exhibits the nucleation nature (Kerner 1998a, b, 1999a, b, c, 2004a, 2009, 2011a, 2013a, 2015b, 2016a, 2017b).

- The empirical nucleation nature of traffic breakdown at a network bottleneck is an empirical fundamental of transportation science.

Generally accepted classical traffic and transportation theories have had a great impact on the understanding of many empirical traffic phenomena (see, e.g., references in reviews and books Chowdhury et al. 2000; Daganzo 1997; Elefteriadou 2014; Gartner et al. 2001; Gazis 2002; Helbing 2001; Kerner 2004a, 2009, 2013a, 2015b, 2016a, 2017b; May 1990; Nagel et al. 2003; Newell 1982; Schadschneider et al. 2011; Treiber and Kesting 2013). Nevertheless, as shown in the book (Kerner 2017b) in detail, the classical traffic and transportation theories have failed by their applications in the real world. Even

several decades of very intensive effort to improve and validate network optimization and control models based on the classical traffic and transportation theories have had no success. Indeed, there can be found no examples where online implementations of the network optimization models based on these classical traffic and transportation theories could reduce congestion in real traffic and transportation networks.

In particular, there are a large number of classical approaches to dynamic traffic assignment and management in traffic networks (see references in the reviews and books (Bell and Iida 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as in Chap. 14 of the book (Kerner 2017b)). However, as it has recently been shown in Kerner 2016b, 2017a, applications of the classical approaches for dynamic traffic assignment in traffic networks deteriorate the traffic system while provoking heavy traffic congestion in urban networks. We can call such a dynamic traffic assignment as standard dynamic traffic assignment. This is because this dynamic traffic assignment is related to the state of the art in traffic and transportation research (see references in the reviews and books (Bell and Iida 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as in conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Hoogendoorn et al. 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Taylor 2002)).

The failure of classical traffic and transportation theories can be explained as follows: The classical traffic and transportation theories are not consistent with the nucleation nature of traffic breakdown. The nucleation nature of empirical traffic breakdown was understood only during the last 20 years (Kerner 1998a, b, 1999a, b, c, 2004a). In contrast, the classical theoretical works, in particular, made by Wardrop (1952), Lighthill and Whitham (1995), and Richards (1956), Herman, Gazis, Montroll, Potts, Rothery, and Chandler (Chandler et al. 1958; Gazis et al.

1959, 1961; Herman et al. 1959), Newell (1961, 1963), Kometani and Sasaki (1958, 1959, 1961), Prigogine (1961), Reuschel (1950), and Pipes (1953) that are the basis for the generally accepted fundamentals and methodologies of traffic and transportation theory, have been introduced in the 1950s–1960s. These and other scientists whose ideas led to the classical fundamentals and methodologies of traffic and transportation theory could not know the nucleation nature of real traffic breakdown at road bottlenecks. Respectively, the author introduced the three-phase traffic theory (Kerner 1998a, b, 1999a, b, c) and the breakdown minimization (BM) principle (Kerner 2011a).

- The three-phase theory is the theoretical fundamental of transportation science that explains the empirical nucleation nature of traffic breakdown.
- The BM principle is the theoretical fundamental of transportation science that permits to maximize the network throughput preventing traffic breakdown in the network.

The classical traffic and transportation theories that failed by their applications in the real world are currently the methodologies of teaching programs in most universities, the subject of publications in most transportation research journals and scientific conferences (see, e.g., references in the reviews and books (Bell and Iida 1997; Bellomo et al. 2002; Brockfeld et al. 2003, 2005; Chowdhury et al. 2000; Daganzo 1997; Elefteriadou 2014; Friesz and Bernstein 2016; Gartner et al. 2001; Gazis 2002; Helbing 2001; Maerivoet and De Moor 2005; Mahmassani 2001; May 1990; Nagatani 2002; Nagel et al. 2003; Newell 1982; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Schadschneider et al. 2011; Sheffi 1984; Treiber and Kesting 2013; Wolf 1999) as well as in conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Fukui et al. 2003; Helbing et al. 2000; Hoogendoorn et al. 2005, 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Schadschneider et al. 2007; Schreckenberg and Wolf 1998; Taylor 2002; Wolf et al. 1995)). This criticism of the author on the traffic and

transportation theories that are related to the state of the art in the traffic and transportation research community has been proven in detail in a recent book (Kerner 2017b). In this book, we have proven that:

- (i) Traffic breakdown at a network bottleneck exhibits the nucleation nature.
- (ii) Applications of classical traffic and transportation theories for reducing traffic congestion in the real world have failed.
- (iii) The three-phase theory is incommensurable with the classical traffic theories.
- (iv) Standard dynamic traffic assignment tends to provoke heavy traffic congestion in a traffic network.
- (v) Applications of the BM principle result in the maximization of the network throughput while ensuring that no breakdown can occur in the network.

It should be noted that the book (Kerner 2017b) is about 650 pages. Therefore, the book is devoted to readers that would like to understand in detail how the author has proven the criticism on classical traffic and transportation theories and their ITS-applications related to the state of the art in the traffic and transportation research. In contrast to the book (Kerner 2017b), this brief entry is devoted to scientists that do not necessarily work in the fields of traffic and transportation science. Thus, rather than a consideration of a diverse variety of empirical spatiotemporal traffic flow phenomena in traffic networks and of traffic flow theories and models that should explain them, we focus on a discussion of the *reason* for the failure of the applications of the classical traffic flow theories and *consequences* of this failure for the future traffic flow organization in traffic networks.

Moreover, in the book (Kerner 2017b) there are more than 1200 references to works of several generations of traffic and transportation researchers. For this reason, in this entry, we limit referring only to some original works as well as to some reviews and books needed for the understanding of results and conclusions of this brief critical review.

To give readers a possibility to make their own impression about the validity of the criticism of the

author on classical traffic and transportation theories and their ITS-applications, in this volume of *Encyclopedia of Complexity*, we have included review articles that describe ITS-applications for traffic management based on the classical traffic and transportation theories (Gartner and Stamatiadis 2009; Hegyi et al. 2017; Rakha and Tawfik 2009). In our entry, while explaining what and why we criticize the generally accepted fundamentals and methodologies of classical traffic and transportation theories, we will often refer to these reviews (Gartner and Stamatiadis 2009; Hegyi et al. 2017; Rakha and Tawfik 2009).

Empirical Nucleation Nature of Traffic Breakdown: Empirical Fundamental of Transportation Science

Definition of Traffic Phases in Three-Phase Traffic Theory

From many empirical studies of real field traffic data measured in different countries over many years of traffic flow observations, it has been

found that in congested traffic two qualitatively different traffic phases should be distinguished: synchronized flow (S) and wide moving jam (J) (Kerner 1998a, b, 1999a, b, c, 2004a). Therefore, in the three-phase traffic flow theory introduced by the author (three-phase theory for short), there are three phases (Kerner 1998a, b, 1999a, b, c, 2004a):

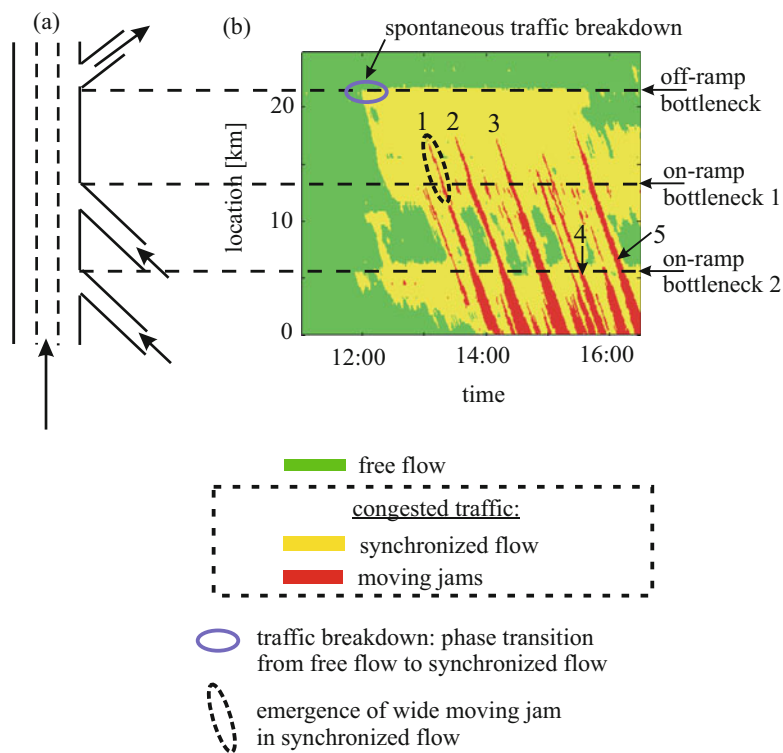
- 1. Free flow (F)
- 2. Synchronized flow (S)
- 3. Wide moving jam (J)

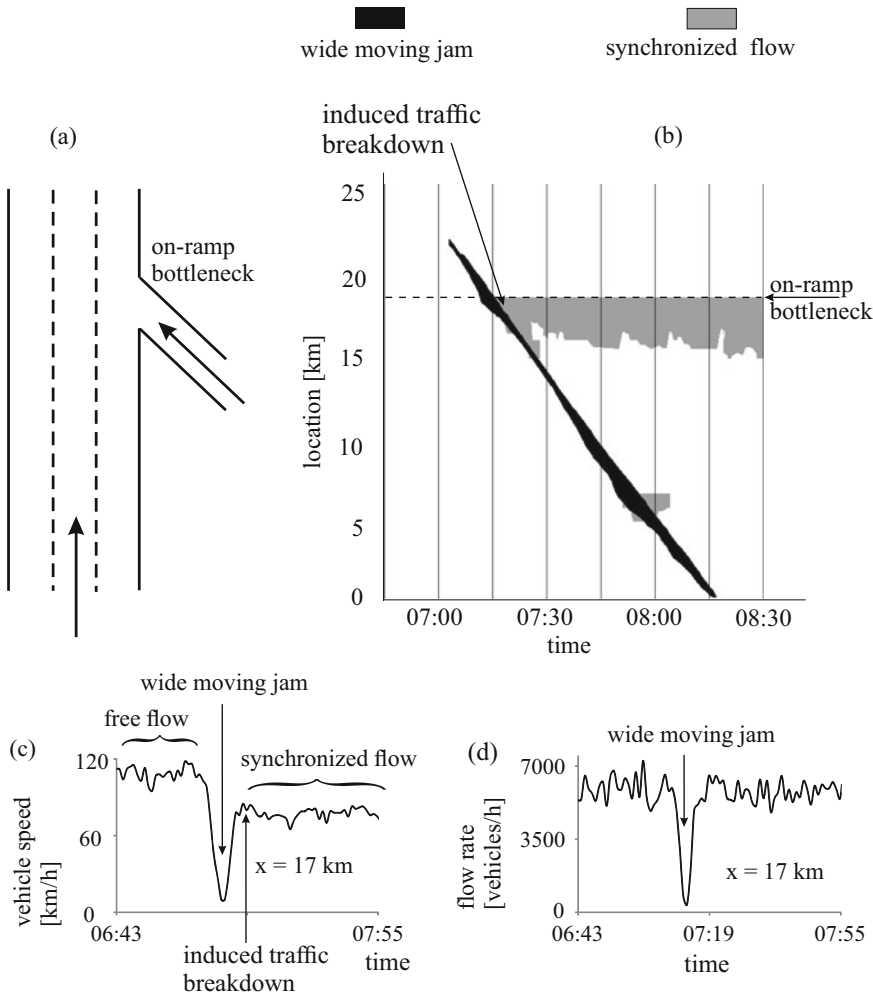
In the three-phase theory, the traffic phases *wide moving jam* and *synchronized flow* in congested traffic are defined, respectively, through the use of the definitions [J] and [S] as follows (Figs. 1 and 2) (Kerner 1998a, b, 1999a, b, c, 2004a):

[J] A wide moving jam is a moving jam that maintains the mean velocity of the downstream front of the jam, even as the jam propagates through other different traffic states of free flow and synchronized flow or highway bottlenecks.

Breakdown in Traffic Networks,

Fig. 1 Empirical example of a spontaneous traffic breakdown with subsequent formation of complex congested traffic pattern measured on a three-lane highway A5-North in Germany on March 23, 2001: (a) Sketch of highway section with three highway bottlenecks (off-ramp bottleneck and two on-ramp bottlenecks). (b) Empirical speed data presented in space and time with some averaging method (Taken from Kerner 2004a)





Breakdown in Traffic Networks, Fig. 2 Example of empirical induced traffic breakdown at on-ramp bottleneck that is induced by wide moving jam propagation: (a) Sketch of section of three-lane highway in Germany with an on-ramp bottleneck. (b) Speed data presented in space and time with some averaging method; data is measured with road detectors installed along road section in (a). (c, d)

Speed (c) and total flow rate (d) over time at the location of the bottleneck $x = 17$ km. Average data of vehicle speed in space and time (1-min averaging interval). Real field traffic data of road detectors measured on three-lane freeway A5-South in Germany on June 23, 1998 (Taken from Kerner 2004a)

This is a characteristic feature [J] of the wide moving jam phase (Kerner and Konhäuser 1994; Kerner and Rehborn 1996a, b, 1997).

[S] In contrast to the wide moving jam traffic phase, the downstream front of the synchronized flow phase does *not* maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often *fixed* at a bottleneck. In other words, the synchronized

flow phase does not show the characteristic feature [J] of the wide moving jam phase.

Moving jams labeled by numbers 1–5 in Fig. 1b propagates through on-ramp bottleneck 1 and on-ramp bottleneck 2 while maintaining the mean velocity of the jam downstream front. Therefore, these moving jams are related to the wide moving jam phase. A moving jam shown in another empirical example (Fig. 2b) propagates

through an on-ramp bottleneck while maintaining the mean velocity of the jam downstream front. Therefore, the moving jam is also related to the wide moving jam phase.

The downstream front of synchronized flow separates synchronized flow upstream of the bottleneck from free flow downstream. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front. In Fig. 1b, we can see congested traffic whose downstream front is fixed at the off-ramp bottleneck (dashed line related to the off-ramp bottleneck in Fig. 1). Thus, congested traffic resulting from traffic breakdown satisfies the definition [S] for the synchronized flow phase. The same conclusion can be made for congested traffic whose downstream front is fixed at the on-ramp bottleneck in Fig. 2b: This congested traffic belongs to the synchronized flow phase.

In the synchronized flow phase (S), the average speed is smaller than the average speed in the free flow phase (F), whereas the flow rate can be as large as the flow rate in the phase F (Fig. 2c). In contrast with the phase S, within the wide moving jam phase (J), both the speed and flow rate are very small, sometimes as small as zero (Fig. 2d) (Additional explanations of the terms “synchronized flow” and “wide moving jam” can be found in Sec. 2.4.4 of the book (Kerner 2009)).

Empirical Traffic Breakdown

Traffic breakdown at a highway bottleneck is a local phase transition from free flow (F) to congested traffic whose downstream front is usually fixed at the bottleneck location (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990). Therefore, highway capacity of free flow is limited by traffic breakdown (see references in Elefteriadou 2014; Elefteriadou et al. 2014; Hall 2001; May 1990). Empirical studies show that during traffic breakdown vehicle speed sharply decreases. In contrast, after traffic breakdown has occurred, the flow rate can remain as large as in an initial free flow. For this reason, traffic breakdown is also called *speed drop*

or *speed breakdown* (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990).

The empirical result that the downstream front of traffic congestion resulting from traffic breakdown is fixed at the bottleneck is a *common result* of all empirical observations of traffic breakdown at any highway bottleneck (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; Kerner 2004a; May 1990). As abovementioned, in the three-phase theory, this congested traffic belongs to the synchronized flow phase (Figs. 1b and 2b). In other words, traffic breakdown is a transition from free flow to synchronized flow (called $F \rightarrow S$ transition) (Kerner 2004a).

- Traffic breakdown at a highway bottleneck is a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition): Traffic breakdown at the highway bottleneck and the $F \rightarrow S$ transition are synonyms.

Empirical Nucleation Nature of Traffic Breakdown

In Kerner (1998a, b, 1999a, b, c, 2004a), it has been revealed that empirical traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck exhibits the nucleation nature. The empirical nucleation nature of the $F \rightarrow S$ transition is as follows:

- Traffic breakdown is a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) that occurs in a metastable state of free flow at a highway bottleneck with respect to the $F \rightarrow S$ transition: Small local speed disturbances in the metastable state of free flow decay. However, when a large enough speed disturbance appears in the metastable state of free flow at the bottleneck, traffic breakdown does occur.
- As in other metastable systems of natural science, there can be spontaneous traffic breakdown at a bottleneck (Fig. 1b) and induced traffic breakdown at a bottleneck (Fig. 2b).

- The evidence of the induced traffic breakdown at a bottleneck in real traffic (see an empirical example in Fig. 2b) is the empirical proof of the nucleation nature of empirical traffic breakdown at a highway bottleneck.

To understand a metastable state of free flow with respect to the $F \rightarrow S$ transition at the bottleneck in more details, we consider a local speed disturbance in free flow that causes a local speed decrease at the bottleneck. The disturbance is localized at the bottleneck. We use the term “the amplitude of the disturbance”: The larger the disturbance amplitude, the smaller the free flow speed within the disturbance in comparison with the free flow speed outside the disturbance. If the amplitude of the disturbance is small enough, then no $F \rightarrow S$ transition occurs at the bottleneck. However, when the amplitude of a local speed disturbance in the metastable free flow is equal to or exceeds a critical amplitude, the $F \rightarrow S$ transition does occur.

A local speed disturbance occurring in the metastable free flow that leads to the $F \rightarrow S$ transition is called a nucleus for the $F \rightarrow S$ transition. Respectively, a local speed disturbance in the metastable free flow with a critical amplitude required for traffic breakdown ($F \rightarrow S$ transition) is called a critical nucleus. In other words, in real traffic there is the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck. In more details, such a macroscopic consideration of a nucleus required for $F \rightarrow S$ transition can be found in Sect. 5.3, whereas a microscopic theory of a nucleus required for $F \rightarrow S$ transition can be found in Sects. 5.13 and 5.13 of the book (Kerner 2017b) (see also an entry Kerner and Klenov (2009) of this Springer Encyclopedia).

Detailed empirical studies of the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway bottlenecks have been made in Chap. 3 of the book (Kerner 2017b).

As abovementioned, the term *traffic breakdown at a highway bottleneck* is a synonym of the term *$F \rightarrow S$ transition at the bottleneck*. Therefore, the term *a nucleus for $F \rightarrow S$ transition at the bottleneck* is a synonym of the term *a nucleus for traffic breakdown at the bottleneck*.

Traffic Breakdown: Basic Empirical Phenomenon of Transportation Science

Vehicular traffic in a traffic network is a spatio-temporal process because it occurs in space and time. Traffic and transportation networks are usually very complex. Therefore, it is not surprising that in empirical studies of traffic data measured in vehicular traffic, a diverse variety of empirical spatiotemporal traffic phenomena have been discovered.

Obviously, each of the traffic and transportation theories and models can explain some real traffic phenomena, and each of the models exhibits a limited region of the applicability for the explanation of real traffic and transportation phenomena. Therefore, the following question arises:

- Is there an empirical traffic phenomenon that can be considered a fundamental empirical basis of transportation science?

This phenomenon should satisfy the following requirement:

- A traffic and/or transportation theory, which is a theoretical basis for reliable dynamic management, control, assignment, and organization of traffic and transportation networks as well as other ITS-applications, must be consistent with the basic empirical features of this phenomenon.

Users of traffic and transportation networks would expect that through the use of traffic control, dynamic traffic assignment and other methods of dynamic optimization traffic breakdown can be prevented, i.e., free flow can be maintained in the network. This is because due to traffic breakdown, congested traffic occurs, in which travel time, fuel consumption, as well as other travel costs increased considerably in comparison with travel costs in free flow.

Usually, after traffic breakdown has occurred at a network bottleneck, the upstream front of the congested pattern propagates upstream. Over time moving jams emerge often within the pattern (Fig. 1b). In this case, the mean flow rate and the average speed within congested traffic can

become considerably smaller than, respectively, the flow rate and the average speed in free flow upstream of the congested pattern.

- Congested traffic, in which the flow rate reduction through congestion is large enough, can be considered *heavy traffic congestion*.

To dissolve heavy congested traffic in urban networks, for example, through the reduction of the flow rate in free flow upstream of a congested traffic pattern, a great limitation of incoming traffic into a congested part of the network through the use of ITS is required. This can lead to heavy traffic congestion in other parts of the network.

Thus, after heavy traffic congestion has already developed in the network (e.g., as shown in Fig. 1b), it is very difficult with the use of ITS to dissolve heavy traffic congestion or even to limit the upstream propagation of traffic congestion in the network. This leads to a well-known conclusion that rather than controlling heavy traffic congestion in the network, ITS should prevent either traffic breakdown or limit the development of traffic congestion in the network.

This explains why the understanding of empirical features of real traffic breakdown has the fundamental priority for the development of reliable ITS. Therefore, any traffic and transportation theory, which is claimed to be a basis for the development of reliable ITS-applications, in particular, for automated driving (called also as self-driving, autonomous driving, or automatic driving), cooperative driving, methods and strategies for traffic control, dynamic traffic assignment, as well as dynamic network optimization, should be consistent with the fundamental empirical features of traffic breakdown at a highway bottleneck. As explained above, the nucleation nature of real traffic breakdown at highway bottlenecks is the fundamental empirical feature of the breakdown. Consequently, we can also make the following conclusion.

- The understanding of empirical nucleation nature of real traffic breakdown at a highway bottleneck has the fundamental priority for the development of reliable ITS.

- The nucleation nature of empirical traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck can be considered an empirical fundamental of transportation science.
- Traffic and transportation theories, which are not consistent with the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway bottlenecks, cannot be applied for the development of reliable traffic control, dynamic traffic assignment, as well as other reliable ITS-applications in traffic and transportation networks.

Possible questions of readers like why the criticism of generally accepted classical traffic and transportation theories can be solely based on an analysis whether the classical traffic and transportation theories are consistent with the empirical nucleation nature of traffic breakdown at a highway bottleneck or not as well as another possible question why traffic breakdown can be considered independent of resulting congested patterns have been explained in Chap. 1 of the book (Kerner 2017b).

Failure of Classical Traffic Flow Models and Theories

Earlier traffic flow models and theories, which claim to explain and/or predict traffic breakdown, have been initiated in the pioneering works by Lighthill and Whitham (1995) as well as by Richards (1956) (Lighthill-Whitham-Richards (LWR) model), Herman, Gazis, Montroll, Potts, Rothery, and Chandler (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959), Kometani and Sasaki (1958, 1959, 1961), Prigogine (1961), Reuschel (1950), Payne (1971, 1979), Pipes (1953), Newell (1961, 2002), Gipps (1981, 1986), Wiedemann (1974), Nagel and Schreckenberg (Barlović et al. 1998; Nagel and Schreckenberg 1992), and by Daganzo (1994, 1995). These classical traffic flow models and theories are the theoretical basis for a huge number of other traffic flow models that can be found, for example, in reviews and books (Bellomo et al. 2002; Chowdhury et al. 2000; Cremer 1979;

Daganzo 1997; Elefteriadou 2014; Gartner et al. 2001; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Maerivoet and De Moor 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Newell 1963; Papageorgiou 1983; Saifuzzaman and Zheng 2014; Schadschneider et al. 2011; Shvetsov 2003; Treiber and Kesting 2013; Whitham 1974; Wolf 1999). These and many other classical traffic flow theories and models have made a great impact on the understanding of many traffic phenomena.

However, as explained in Chaps. 4 and 8 of the book (Kerner 2017b) in detail, all earlier traffic flow models and theories have failed in the explanation and/or simulations of the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck: None of these models and theories is able to show and predict the metastability of free flow with respect to an $F \rightarrow S$ transition at the bottleneck. A detailed consideration of these critical conclusions is also the subject of this volume of the Encyclopedia (Traffic Breakdown, Modeling Approaches to) (Kerner 2017c).

- The classical traffic flow theories have failed to show the empirical fundamental of transportation science – the nucleation nature of traffic breakdown at a bottleneck with respect to an $F \rightarrow S$ transition in metastable free flow at the bottleneck.
- Therefore, the classical traffic flow theories cannot be applied for the development of reliable ITS-applications for dynamic traffic control and management in traffic networks.

Theoretical Fundamental of Transportation Science: The Three-Phase Theory

The three-phase theory introduced by the author results from a spatiotemporal analysis of real field traffic data measured during many years of traffic observations on different highways in a variety of countries. The three-phase theory formulated first in 1998–1999 (Kerner 1998a, b, 1999a, b, c) explains features of the three traffic phases – free

flow, synchronized flow, and wide moving jam – as well as features of phase transitions between the traffic phases.

- The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown at highway bottlenecks.

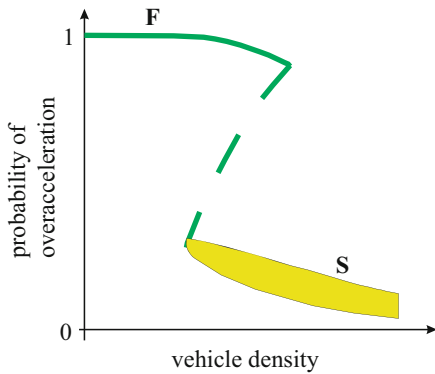
Therefore, in this entry we consider briefly only the explanation of traffic breakdown at a highway bottleneck with the three-phase theory. A detailed consideration of traffic breakdown at the bottleneck can be found in an entry of Springer Encyclopedia (Kerner 2017c). A detailed consideration of the three-phase theory can be found in the books (Kerner 2004a, 2009, 2017b).

$S \rightarrow F$ Instability and Z-Characteristic for Traffic Breakdown

The empirical nucleation nature of traffic breakdown at a highway bottleneck results from the metastability of free flow with respect to an $F \rightarrow S$ transition at the bottleneck (section “Empirical Nucleation Nature of Traffic Breakdown”).

In the three-phase theory, the metastability of free flow with respect to an $F \rightarrow S$ transition has been explained through the following hypotheses (Kerner 2004a):

- The metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck is caused by a competition between two opposing tendencies – the tendency toward free flow due to the *over-acceleration effect* (driver over-acceleration) and the tendency toward synchronized flow due to the *speed adaptation effect* (driver speed adaptation).
- The tendency toward free flow due to the over-acceleration effect is associated with a traffic flow instability in synchronized flow caused by a discontinuous character of a density (flow rate) function of the probability of driver over-acceleration. The discontinuous character of driver over-acceleration is associated with the existence of a driver time delay in over-acceleration (Fig. 3).
- There is an instability in synchronized flow that causes a growing speed wave of a local increase



Breakdown in Traffic Networks, Fig. 3 Qualitative illustration of discontinuous character of density dependence of the probability of driver over-acceleration in three-phase theory (Kerner 1999a, b, c, 2004a). F – free flow, S – synchronized flow. Dashed curve between F and S states shows qualitatively unstable part of density dependence of the probability of driver over-acceleration

in the speed in synchronized flow. The growth of the speed wave leads to a phase transition from synchronized flow (S) to free flow (F) ($S \rightarrow F$ transition). For this reason, we call this instability as an $S \rightarrow F$ instability (Fig. 4). The $S \rightarrow F$ instability results from the discontinuous character of driver over-acceleration.

A brief explanation of these hypotheses of the three-phase theory is as follows (detailed explanations can be found in Chap. 5 of the book (Kerner 2017b)).

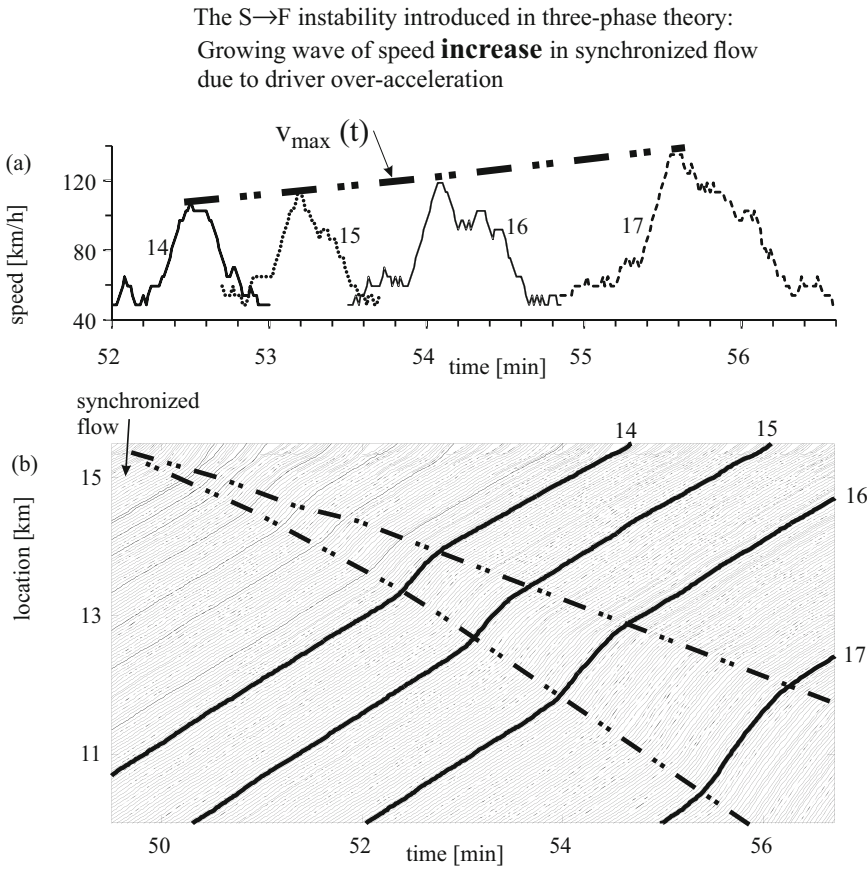
The driver speed adaptation effect occurs within so-called synchronization space gap. If a driver approaches a slower moving preceding vehicle that the driver cannot overtake, the driver adapts the vehicle speed to the speed of the preceding vehicle, when the space gap to the preceding vehicle is smaller than the synchronization space gap. At each vehicle speed within synchronized flow, while adapting the vehicle speed to the speed of the preceding vehicle within the space gap range between the synchronization space gap and a safe space gap, a driver can make an arbitrary choice in the space gap: The driver accepts different space gaps at different times and does not control a fixed space gap to the preceding vehicle.

The $S \rightarrow F$ instability introduced in three-phase theory: Growing wave of speed **increase** in synchronized flow due to driver over-acceleration.

Contrarily to the speed adaptation effect, the over-acceleration effect is a driver maneuver leading to a higher speed from initial car following at a lower speed. It is assumed in the three-phase theory (Kerner 2004a, 2009) that due to a driver time delay in over-acceleration, the probability of driver over-acceleration in car following exhibits a discontinuous character. The discontinuous character of the over-acceleration leads to the $S \rightarrow F$ instability, i.e., to the emergence of a growing wave of a local increase in the speed in synchronized flow. This traffic flow instability ($S \rightarrow F$ instability) can be considered the result of the over-acceleration effect that determines the tendency toward free flow within a local speed increase in synchronized flow.

A microscopic theory of the $S \rightarrow F$ instability has been developed only recently (Kerner 2015a). In this theory, it has been proven that the $S \rightarrow F$ instability exhibits the nucleation nature. If a large enough local speed increase appears randomly in metastable synchronized flow with respect to the $S \rightarrow F$ instability, an $S \rightarrow F$ instability is realized. Such a local speed increase can be considered a nucleus for the $S \rightarrow F$ instability. Otherwise, a small enough local speed increase in the metastable synchronized flow decays over time. In the three-phase theory, it has been found that the nucleation nature of the $S \rightarrow F$ instability governs the nucleation nature of the $F \rightarrow S$ transition (traffic breakdown) at a highway bottleneck (a detailed consideration of the $S \rightarrow F$ instability can be found in Secs. 5.12 and 5.13 of the book (Kerner 2017b)).

We can state that a local speed disturbance in free flow at a highway bottleneck becomes a nucleus required for traffic breakdown, if the over-acceleration effect within this speed disturbance is weaker than the speed adaptation effect. In this case, no $S \rightarrow F$ instability occurs. Therefore, the speed adaptation effect results in traffic breakdown at the bottleneck. Otherwise, if within a local speed disturbance, the over-acceleration effect is stronger than the speed adaptation effect; an $S \rightarrow F$ instability occurs within the disturbance. Thus, the $S \rightarrow F$ instability prevents traffic



Breakdown in Traffic Networks, Fig. 4 Simulations of $S \rightarrow F$ instability introduced in three-phase theory: (a, b) Growing wave of speed increase in synchronized flow due

to driver over-acceleration as time function (a) and on vehicle trajectories (b) (Adapted from Kerner 2015a)

breakdown, and therefore, the initial local speed disturbance decays over time.

A simple representation of the metastability of traffic breakdown at the bottleneck is given by a Z-characteristic for the $F \rightarrow S$ transition (traffic breakdown) (Fig. 5). The Z-characteristic results from the competition between the speed adaptation and over-acceleration effects in free flow at a highway bottleneck. The Z-characteristic is also called the Z-characteristic for traffic breakdown. Explanations of the Z-characteristic for traffic breakdown have been made in the entry of this volume (Kerner 2017c) as well as in Chap. 5 of the book (Kerner 2017b).

Minimum and Maximum Highway Capacities

With the use of the Z-characteristic for traffic breakdown, the empirical nucleation nature of

traffic breakdown at the bottleneck is explained as follows. The metastability of free flow with respect to the $F \rightarrow S$ transition is realized within a flow rate range (Kerner 2004a)

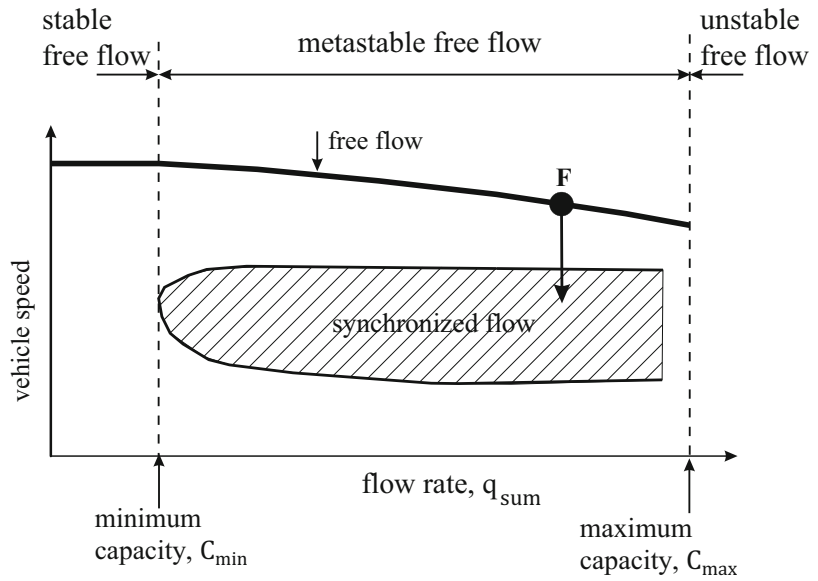
$$C_{\min} \leq q_{\text{sum}} < C_{\max}, \quad (1)$$

where q_{sum} is the flow rate in free flow at a highway bottleneck, $C_{\min}$ is a minimum highway capacity, and $C_{\max}$ is a maximum highway capacity. Because under conditions (1) free flow is in a metastable state with respect to the $F \rightarrow S$ transition, traffic breakdown can be induced at the bottleneck at any flow rate q_{sum} satisfying (1). This leads to the following conclusions (Kerner 2004a):

- *At any time instant*, there is an infinite number of the flow rates q_{sum} (1) in free flow at a

Breakdown in Traffic Networks,

Fig. 5 Qualitative explanation of the empirical metastability of free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at bottleneck (Kerner 2004a; Kerner and Klenov 2003). Qualitative Z-speed-flow rate characteristic for traffic breakdown; F – free flow, S – synchronized flow (two-dimensional (2D) hatched region) (results of simulations of the Z-speed-flow rate characteristic for traffic breakdown (Kerner and Klenov 2003) are discussed in Kerner 2017c)



highway bottleneck at which traffic breakdown can be induced at the bottleneck. These flow rates are highway capacities of free flow at the bottleneck.

- At any time instant, there is an infinite number of highway capacities C of free flow at the bottleneck. The range of the infinite number of highway capacities is limited by the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$ (Fig. 5):

$$C_{\min} \leq C \leq C_{\max}, \quad (2)$$

where $C_{\min} < C_{\max}$. The existence of an infinite number of highway capacities at any time instant means that highway capacity is stochastic.

The physical sense of this capacity definition is as follows (Fig. 5). Highway capacity is limited by traffic breakdown in an initial free flow at a highway bottleneck. In other words, any flow rate q_{sum} in free flow at the bottleneck at which traffic breakdown can occur is highway capacity. At any time instant, there is an infinite number of such highway capacities $C = q_{\text{sum}}$ at which traffic breakdown can occur. These capacities satisfy conditions (2). Thus, in the three-phase theory, any flow rate q_{sum} in free flow at a highway bottleneck that satisfies conditions.

$$C_{\min} \leq q_{\text{sum}} \leq C_{\max} \quad (3)$$

is equal to one of the stochastic highway capacities of free flow at the bottleneck (Fig. 5).

Under condition (1), free flow is in a *metastable* state with respect to an $F \rightarrow S$ transition (traffic breakdown). When condition.

$$q_{\text{sum}} = C_{\max} \quad (4)$$

is satisfied, we assume that the probability of traffic breakdown during a given time interval is equal to 1 (see Sect. 5.4.8 of the book (Kerner 2017b)). For this reason, free flow can be considered as an *unstable* one with respect to traffic breakdown at the bottleneck. This explains the difference between conditions (1) that include only metastable states of free flow and conditions (3) that include both metastable states of free flow (1) and the unstable state of free flow related to condition (4).

Figure 5 together with conditions (1) explains the term *metastable free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at the bottleneck* as follows (labeled by “metastable free flow” in Fig. 5). For each flow rate q_{sum} in free flow at the bottleneck that satisfies conditions (1), there can be a transition from a state of the free

flow traffic phase (F) to a state of the synchronized flow traffic phase (S) at the bottleneck. This $F \rightarrow S$ transition (traffic breakdown) at the bottleneck occurs only, when a nucleus required for the breakdown appears at the bottleneck. An arrow $F \rightarrow S$ shown in Fig. 5 illustrates symbolically one of possible $F \rightarrow S$ transitions occurring in a metastable state of free flow when the nucleus appears in the free flow at the bottleneck.

When condition

$$q_{\text{sum}} < C_{\text{min}} \quad (5)$$

is satisfied, no traffic breakdown can occur. This is independent of the amplitude of a time-limited local disturbance in free flow at the bottleneck. Indeed, we assume that under condition (5) free flow is *stable* with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck (labeled by “stable free flow” in Fig. 5).

Respectively, when condition.

$$q_{\text{sum}} \geq C_{\text{max}} \quad (6)$$

is satisfied, traffic breakdown does occur at the bottleneck during a given time interval (see Sect. 5.4.8 of the book (Kerner 2017b)). This is because we assume that under condition (6) free flow is *unstable* with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck (labeled by “unstable free flow” in Fig. 5).

Incommensurability of Classical Traffic Theories with Three-Phase Theory

In Sects. 1.13 and 1.14 of the book (Kerner 2017b), we have stressed that rather than empirical features of congested traffic patterns, the metastability of free flow at a highway bottleneck with respect to an $F \rightarrow S$ transition (traffic breakdown) is needed to use as the basic criterion for a comparison of the classical traffic flow theories with the three-phase theory. This is because none of the classical traffic flow theories and models incorporates the metastability of free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at a highway bottleneck.

- The metastability of free flow with respect to the $F \rightarrow S$ transition (traffic breakdown) at the bottleneck has no sense for the classical traffic theories.

The metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck is the main feature of the three-phase theory. This shows the incommensurability of the three-phase theory and the classical traffic flow theories. The term *incommensurability* has been introduced by Kuhn in his book (Kuhn 2012) to explain a paradigm shift in a scientific field. The incommensurability of the three-phase theory and the classical traffic flow theories as well as its consequences for transportation science has been considered in detail in Chap. 8 of the book (Kerner 2017b).

Classical Understanding of Stochastic Highway Capacity

Traffic breakdown at a highway bottleneck limits highway capacity (see, e.g., Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Ceder 1999; Daganzo 1997; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Greenshields et al. 1935; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; Highway Capacity Manual 2000, 2010; Lesort 1996; Lorenz and Elefteriadou 2000; Mahmassani 2005; May 1990; Persaud et al. 1998; Taylor 2002; Zurlinden 2003). This means that the origin of traffic breakdown at the bottleneck determines the physical nature of highway capacity. Beginning from the classical work by Greenshields (Greenshields et al. 1935), there have been a huge number of studies of highway capacity (see, e.g., Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Ceder 1999; Daganzo 1997; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Greenshields et al. 1935; Hall 1987, 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; Highway Capacity Manual 2000, 2010; Lesort 1996; Lorenz and Elefteriadou 2000; Mahmassani 2005; May 1990; Persaud et al. 1998; Taylor 2002; Zurlinden 2003).

In 1995 Elefteriadou et al. (1995) found that traffic breakdown exhibits a probabilistic

(stochastic) nature. The empirical investigations of traffic breakdown have led to the following definition of stochastic highway capacity (see, e.g., Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 2014).

- *At any given time instant*, there is a particular value of stochastic highway capacity of free flow at a bottleneck.

Because empirical highway capacity is a stochastic value, we do not know and we cannot measure this particular (i.e., a single) value of highway capacity of free flow at a bottleneck at a given time instant. However, in accordance with the above classical understanding of highway capacity, it is assumed that at *any given time instant*, there is a particular value of stochastic highway capacity of free flow at a bottleneck (see, e.g., Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 2014 and references there). Because this understanding of highway capacity is accepted in the transportation research community (see, e.g., the book by Elefteriadou (2014) as well as a paper Elefteriadou et al. (2014) and references there), we call it as a classical understanding of stochastic highway capacity (or stochastic highway capacity in the classical theory).

As abovementioned, the empirical nucleation nature of traffic breakdown at a highway bottleneck is an empirical fundamental of transportation science. However, as explained in Sect. 4.10 of the book (Kerner 2017b), the classical understanding of stochastic highway capacity contradicts this empirical fundamental of transportation science:

- The classical understanding of stochastic highway capacity is inconsistent with the empirical nucleation nature of traffic breakdown at a highway bottleneck.

The understanding of stochastic highway capacity that is consistent with the empirical nucleation nature of traffic breakdown at a highway bottleneck has been given in the three-phase theory (section “[Minimum and Maximum Highway Capacities](#)”):

- In the three-phase theory, *at any time instant*, there is an infinite number of highway capacities C of free flow at the bottleneck that are between the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$ (Fig. 5).

Thus, from the empirical nucleation nature of traffic breakdown at a highway bottleneck follows that *at any time instant*, there is an infinite number of highway capacities C satisfying conditions $C_{\min} \leq C < C_{\max}$. Contrarily, in the classical understanding of stochastic highway capacity, it is assumed that *at any time instant*, there is a particular value of highway capacity that we do not know due to the stochastic nature of capacity.

As mentioned, the classical understanding of stochastic highway capacity is invalid for real traffic (see Sect. 4.10 of the book Kerner 2017b): The classical understanding of stochastic highway capacity is inconsistent with the empirical nucleation nature of traffic breakdown at a highway bottleneck.

The theoretical basic of the most of ITS-applications based on classical traffic flow theories is the assumption that for any time instant, there is a particular (single) value of highway capacity. This is one of the main reasons for the failure of ITS-applications based on the classical traffic flow theories that we will briefly discuss in section “[Failure of Intelligent Transportation Systems \(ITS\) Based on Classical Traffic Theories](#).”

Failure of Intelligent Transportation Systems (ITS) Based on Classical Traffic Theories

As abovementioned, the classical traffic and transportation theories have made a great impact on the understanding of many traffic phenomena. However, network optimization approaches based on these theories have failed by their applications in the real world. Traffic engineers that work with real installations for traffic control know this fact: There can be found no examples where online implementations of the network optimization models based on the classical traffic theories

could reduce congestion in real traffic and transportation networks. It is understandable that traffic researchers would like to publish positive results of applications of their traffic control approaches. This can explain why it is impossible to find publications in which the failure of traffic applications based on the classical traffic and transportation theories in real traffic installations is clearly admitted.

However, in 1995 when the author gave a talk “traffic moving jam without obvious reason” (Kerner and Konhäuser 1994) at the traffic engineering consulting firm Heusch/Boesefeldt GmbH in Aachen (Germany), the failure of traffic applications of the classical theories in real traffic installations was openly deplored by the engineers. This encounter with the traffic engineers of the Heusch/Boesefeldt GmbH has affected the author to begin with a study of real field traffic data (see references in an entry by Rehborn and Klenov (2009) as well as references in an entry of Rehborn et al. (2017) in this volume of the Encyclopedia (Prediction of Traffic Congestion)).

While working during many years in many engineering projects in the field “traffic” at the Daimler Company, the author recognized that no noticeable progress was done in the field of applications of classical traffic and transportation theories for the reduction of traffic congestion in real traffic and transportation networks.

A Myth About “Optimal” Occupancy and “Optimal” Flow Rate

To illustrate the reason for the failure of ITS-applications based on classical traffic theories, we consider a theoretical fundamental diagram of traffic flow (Fig. 6a). The fundamental diagram reflects the obvious empirical feature of traffic flow: The larger the vehicle density, the smaller the average vehicle speed. Thus, a flow rate–density relationship exhibits a maximum point (Fig. 6a). In many ITS-applications based on the classical traffic flow theories, it is assumed that highway capacity is equal to the flow rate at the maximum point of the fundamental diagram denoted in Fig. 6 by $q_{\text{sum}} = q_{\text{cap}}$ (see, e.g., reviews Hegyi et al. 2017; May 1990; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000,

2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003). Respectively, the vehicle density at the maximum point at the fundamental diagram is the critical density $\rho = \rho_{\text{cr}}$ in free flow related to highway capacity (Fig. 6a). Left to the maximum point on the theoretical fundamental diagram in the flow–density plane (Fig. 6a), i.e., when the density $\rho < \rho_{\text{cr}}$, free flow is *stable* (labeled by “F” in Fig. 6). In this case, it is assumed in the classical traffic flow theories that no traffic breakdown can occur at a bottleneck. In contrast, under condition $\rho > \rho_{\text{cr}}$ traffic breakdown should occur leading to congested traffic at the bottleneck (Hegyi et al. 2017; May 1990; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003).

In many traffic feedback control approaches, traffic variables for feedback control are measured through the use of road detectors. However, rather than the density, another traffic variable called *occupancy* o that a road detector can directly measure is usually used in traffic control (see, e.g., reviews Hegyi et al. 2017; May 1990; Papageorgiou et al. 1997, 2003; Papageorgiou and Kotsialos 2000, 2002; Smaragdis and Papageorgiou 2003). The occupancy is determined through formula $o = (T_{\text{veh}}/T_{\text{av}})100\%$, where T_{veh} is the sum of the time intervals within which vehicles are at the detector location; T_{av} is a time interval of traffic variable averaging (often $T_{\text{av}} = 1\text{min}$). The theoretical flow–occupancy relationship (Fig. 6b) is qualitatively the same as that of the theoretical fundamental diagram (Fig. 6a). For this reason, it is assumed that under condition

$$o < o_{\text{cr}} \quad (7)$$

free flow is stable, i.e., no traffic breakdown can occur at the bottleneck in the framework of classical traffic flow theories (Hegyi et al. 2017; May 1990; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003). In (7), the occupancy $o = o_{\text{cr}}$ is the critical occupancy related to the

maximum point at the theoretical flow–occupancy relationship at which the flow rate reaches highway capacity $q_{\text{sum}} = q_{\text{cap}}$ (Fig. 6b).

Therefore, it is assumed in classical traffic flow theories and associated control methods (see, e.g., reviews Hegyi et al. 2017; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) that when a feedback control method maintains the occupancy near some “optimal” occupancy $o_{\text{sum}}^{(\text{opt})}$ that does not exceed the critical occupancy $o = o_{\text{cr}}$, stable free flow should occur at the bottleneck (labeled by “F” in Fig. 6b).

Respectively, in some other traffic control methods based on measurements of the flow rate, it is assumed that stable free flow should persist at the bottleneck when a feedback control method maintains the flow rate downstream of the bottleneck near some “optimal” flow rate in free flow that should not reach the highway capacity $q_{\text{sum}} = q_{\text{cap}}$ (left from the point $q_{\text{sum}} = q_{\text{cap}}$ in Fig. 6) (see the review (Smaragdis and Papageorgiou 2003) and references there).

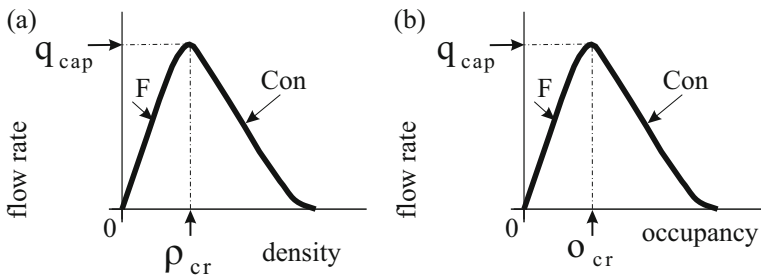
One of the most prominent examples of such ITS-application is ALINEA on-ramp metering of Papageorgiou et al. (see references in Hegyi et al. 2017; Papageorgiou et al. 1990a, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003). ALINEA is a feedback regulator based on the classical control theory. The aim of ALINEA is to maintain the real value of the occupancy o measured by feedback

control detectors near a chosen “optimal” occupancy $o_{\text{sum}}^{(\text{opt})}$ that should not exceed the critical occupancy o_{cr} (Fig. 6b) (Papageorgiou et al. 2007). In accordance with the assumption of Hegyi et al. (2017), Papageorgiou et al. (1997, 2003, 2007), Papageorgiou and Kotsialos (2000, 2002), Papageorgiou and Papamichail (2008), Smaragdis and Papageorgiou (2003), such an application of the classical control theory should maintain a stable state of free flow at an on-ramp bottleneck.

In contrast with this assumption of classical traffic theories, rather than stable free flow (Fig. 6b), free flow is in a *metastable* state with respect to traffic breakdown at the on-ramp bottleneck within the flow rate range between the minimum highway capacity and maximum highway capacity (1) as shown in Fig. 5. For this reason, formula (7) for stable free flow of the classical theories (Daganzo 1997; Hegyi et al. 2017; May 1990; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) is invalid for real traffic.

Indeed, real free flow at the bottleneck is in a metastable state with respect to an $F \rightarrow S$ transition (traffic breakdown) at the bottleneck, when conditions (1) are satisfied. In accordance with conditions (1), for any optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ in free flow within the range

$$o_{\text{min}} \leq o_{\text{sum}}^{(\text{opt})} < o_{\text{cr}}, \quad (8)$$



Breakdown in Traffic Networks, Fig. 6 Theoretical fundamental diagram of traffic flow (see, e.g., reviews Hegyi et al. 2017; May 1990; Papageorgiou et al. 1997, 2003; Papageorgiou and Kotsialos 2000, 2002; Smaragdis

and Papageorgiou 2003): (a) A qualitative flow–density relationship. (b) A qualitative flow–occupancy relationship. F – free flow, Con – congested traffic

free flow is metastable with respect to traffic breakdown at the bottleneck. In (8), the occupancy $o_{\min}$ is related to the flow rate in free flow at the bottleneck $q_{\text{sum}} = C_{\min}$, and the critical occupancy o_{cr} has the same sense as that in the classical traffic flow theories (Fig. 6b) (Hegyi et al. 2017; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003): When $o > o_{\text{cr}}$, in the three-phase theory, free flow is unstable with respect to traffic breakdown as it is assumed in the classical theories. The only difference is that in the three-phase theory, the critical occupancy is related to the flow rate in free flow at the bottleneck $q_{\text{sum}} = C_{\max}$. Thus, at any choice of the optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ that satisfies (8), traffic breakdown can occur in free flow at the bottleneck during a chosen observation time interval with some probability that is larger than zero.

We can make the following conclusions:

- In ITS-applications based on the classical traffic flow theories, it is assumed that there is *critical* occupancy that separates stable free flow from unstable free flow at a highway bottleneck (see, e.g., reviews (Hegyi et al. 2017; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) and references there). This basic assumption of ITS-applications based on the classical traffic flow theories contradicts basically the empirical fundamental of transportation science – the empirical nucleation nature of traffic breakdown at a highway bottleneck.
- Contrarily to ITS-applications based on the classical traffic flow theories, due to the metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck, there is neither a particular optimal flow rate, nor an optimal density, nor else an optimal occupancy in free flow that can be chosen as a target value for a controller of a traffic control method. At any density, any occupancy, or any flow rate related to the range of the flow rate in free flow downstream of the bottleneck (1), traffic breakdown can occur with some finite probability at the bottleneck.
- These conclusions explain why we state that no reliable traffic control and other ITS-applications can be made with the use of the classical traffic flow theories.

A detailed consideration of this criticism can be found in Sects. 10.6.5 and 10.7 of the book (Kerner 2009) as well as in an entry (Kerner 2017c) of this Encyclopedia.

Invalid Simulations of Performance of ITS-Applications

The reason for the failure of ITS-applications based on classical traffic and transportation theories has been explained above: The classical traffic and transportation theories and models cannot explain or they ignore the nucleation nature of traffic breakdown at network bottlenecks observed in real field traffic data.

This criticism is related to all ITS that affect traffic flow, for example, on-ramp metering, variable speed limit control, the effect of cooperative driving with the use of vehicle-to-vehicle (V2 V) communication on traffic flow characteristics, as well as simulations of the performance of dynamic traffic assignment and optimization in traffic and transportation networks, studies of the effect of adaptive cruise control (ACC), automated driving vehicles (automated driving is also called as autonomous driving, automatic driving, or self-driving), and other vehicle systems on traffic flow reviewed, e.g., in Brockfeld et al. (2003, 2005), Daganzo (1997), Elefteriadou (2014), Hegyi et al. (2017), Helbing (2001), May (1990), Papageorgiou et al. (1997, 2003, 2007), Papageorgiou and Kotsialos (2000, 2002), Papageorgiou and Papamichail (2008), Rakha and Tawfik (2009), Saifuzzaman and Zheng (2014), Smaragdis and Papageorgiou (2003), and Treiber and Kesting (2013). In particular, because the classical traffic flow models cannot show the empirical features of the metastability of free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at a highway bottleneck, the application of generally accepted models and associated

simulation tools for a study of the effect of automated driving vehicles on traffic flow can lead to incorrect conclusions. For this reason, such simulations cannot also be used for the development of reliable systems for automated driving vehicles. This criticism is also related to the use of well-known simulation tools based on the classical traffic flow theories like simulation tools VISSIM (Wiedemann model) (Wiedemann 1974) and SUMO (Krauß model) (Krauß et al. 1997). Other references to invalid simulations of ITS-applications based on the classical traffic flow theories can be found in Sects. 1.7.4 and 4.9 of the book (Kerner 2017b).

This explains why applications of the classical traffic flow models for the development of different traffic simulation tools for the analysis of ITS as well as a variety of diverse ITS-applications in which dynamic traffic assignment and control are used together with traffic flow simulations in frameworks of the classical traffic flow models *cannot* be used for reliable and valid analysis of ITS performance in the real world.

Field Trials of On-Ramp Metering

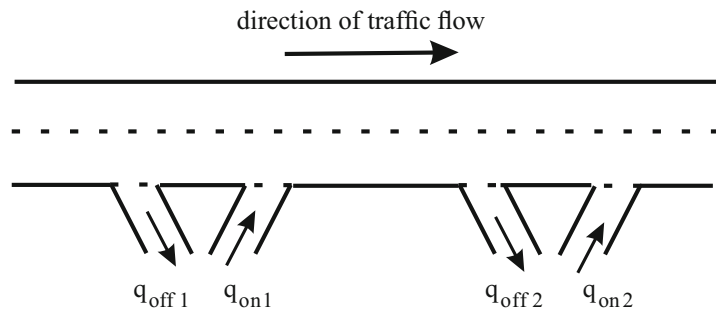
In the entry by Hegyi et al. (2017), field trials of applications of on-ramp metering have been mentioned (Cambridge Systematics, Inc 2001; Levinson and Zhang 2006; Papageorgiou et al. 1990b, 1997) that should prove the reduction of traffic congestion in real traffic and transportation networks. Indeed, in Cambridge Systematics, Inc (2001), Levinson and Zhang (2006), and Papageorgiou et al. (1990b, 1997) have stated that on-ramp metering can reduce or even prevent traffic congestion on some roads of a traffic network. On the first glance, there is a basic

inconsistency between the statement of our entry that “no noticeable progress was done in the field of applications of classical traffic and transportation theories for the reduction of traffic congestion in real traffic and transportation networks” and results of Cambridge Systematics, Inc (2001), Levinson and Zhang (2006), and Papageorgiou et al. (1990b, 1997) related to field trials of on-ramp metering. This inconsistency is resolved as follows: If the on-ramp metering model is inconsistent with the empirical nucleation nature of traffic breakdown, then this reduction or prevention of traffic congestion on some of the network roads can be achieved only at the expense of the occurrence of heavy traffic congestion on other network roads.

For a qualitative explanation of the latter statement (Simulation results that prove this statement can be found in entry Kerner 2017c of this Encyclopedia.), we note that traffic flow occurs on a main highway of an urban network through vehicles merging from on-ramps (two of these on-ramps with on-ramp inflow rates $q_{on\ 1}$ and $q_{on\ 2}$ are shown in Fig. 7). In other words, the flow rate on the highway is mostly determined by on-ramp inflows and off-ramp outflows. Thus, through the use of on-ramp metering one can always reduce the on-ramp inflows to so small values at which the flow rate on the highway is small enough, and therefore, no congestion can occur on the highway. This conclusion is *independent* of the on-ramp metering model used for metering.

However, such a “congestion reduction” on the highway is realized at the expense of those drivers that must wait within a vehicle queue at traffic signals that reduce on-ramp inflows: The latter

Breakdown in Traffic Networks, Fig. 7 A qualitative scheme of a section of two-lane highway with two off-ramps and two on-ramps



drivers should wait in the queue at on-ramps to ensure free flow for other drivers moving on the main highway. This is not acceptable by traffic authorities. In other words, only when the queues at on-ramps are very short and they exist during short enough time intervals, on-ramp metering could be accepted by traffic authorities.

For this reason, it could be clear why the fact whether a model for on-ramp metering is consistent with the empirical fundamental of transportation science or it is not is extremely important for the length and duration of the vehicle queue at traffic signals of on-ramps. This statement has been proven in simulations of a comparison of ALINEA on-ramp metering based on classical traffic theories (section “A Myth About ‘Optimal’ Occupancy and ‘Optimal’ Flow Rate”) and ANCONA on-ramp metering based on the three-phase theory (Kerner 2004a, 2005, 2007a, b, c, 2009) (see Sect. 10.7 of the book (Kerner 2009) as well as the entry (Kerner 2017c) of this Encyclopedia).

Models of on-ramp metering discussed in the reviews (Hegyi et al. 2017; Papageorgiou et al. 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) and empirical papers (Cambridge Systematics, Inc 2001; Levinson and Zhang 2006; Papageorgiou et al. 1990b, 1997) are based on the classical traffic flow theories that are not consistent with the empirical fundamental of transportation science (The exclusion is ANCONA model (Kerner 2004a, 2005, 2007a, b, c, 2009) based on the three-phase theory that has been mentioned in the review (Hegyi et al. 2017).). Because all on-ramp metering models (including the models tested in Cambridge Systematics, Inc 2001; Levinson and Zhang 2006; Papageorgiou et al. 1990b, 1997) are inconsistent with the nucleation nature of traffic breakdown, long queues occur at the signals. For this reason, although there have been very many field trials during the last 40 years of ALINEA (Papageorgiou et al. 1990b, 1997), UP-ALINEA, ALINEA/Q, models applied in Refs. (Cambridge Systematics, Inc 2001; Levinson and Zhang 2006), etc., there are no online business operations of on-ramp metering. Different members of traffic authorities in the

USA and in Germany have explained me that after a short time of such on-ramp metering application of any on-ramp metering models based on classical traffic theories, due to too long (and, therefore, not acceptable) queues occurring on on-ramps, the metering based on these metering models was turned off. This is the result of the inconsistency of real traffic (the empirical nucleation nature of traffic breakdown) and on-ramp metering models based on the classical traffic theories.

In the book Kerner (2017b), readers can find a list of references to such invalid ITS-applications based on classical traffic theories and models as well as explanations of this criticism (see Sects. 1.7.4 and 9.4 of Kerner 2017b). A more detailed consideration of this critical statement for on-ramp metering has been made in Sect. 10.6.5 of the book (Kerner 2009) as well as in entry (Kerner 2017c) in this Encyclopedia.

Thus, we can make the following conclusions to section “Failure of Intelligent Transportation Systems (ITS) Based on Classical Traffic Theories”:

- The reason for the failure of traffic control methods based on classical traffic theories is associated with the nucleation nature of traffic breakdown at network bottlenecks: The traffic control methods are not consistent with this empirical fundamental of transportation science.
- To study the performance of ITS with a traffic flow model, the model should simulate conditions at which the application of ITS can prevent traffic breakdown in a traffic network. However, none of the classical traffic flow models can simulate the empirical nucleation nature of real traffic breakdown at the bottleneck. Therefore, the classical traffic flow models (two-phase models) that have been used and considered, for example, in Brockfeld et al. (2003, 2005), Daganzo (1997), Elefteriadou (2014), Hegyi et al. (2017), Helbing (2001), Papageorgiou et al. (1997, 2003, 2007), Papageorgiou and Kotsialos (2000, 2002), Papageorgiou and Papamichail (2008), Rakha and Tawfik (2009), Saifuzzaman and Zheng (2014), Smaragdis

and Papageorgiou (2003), and Treiber and Kesting (2013), cannot be used for a reliable analysis of ITS performance.

- The three-phase theory is devoted to the explanation of the empirical nucleation nature of real traffic breakdown at the bottleneck. Therefore, only three-phase traffic flow models can be used for a reliable study of ITS-applications.

Congested Pattern Control Approach Based on Three-Phase Theory

When the flow rate at a bottleneck satisfies conditions (1), free flow is in a metastable state with respect to an $F \rightarrow S$ transition (traffic breakdown). Therefore, traffic breakdown can occur regardless of traffic control. This empirical evidence has been the reason for a congested pattern control approach introduced by the author Kerner (2004a, 2005, 2007a, b, c, 2009).

In congested pattern control approach, *no control* of traffic flow at a network bottleneck is realized as long as free flow is realized at the bottleneck. This means that the occurrence of random traffic breakdown is permitted to occur at the bottleneck: Congestion at the bottleneck is allowed to set in. The main idea of this approach is as follows. Only after traffic breakdown has occurred, traffic control starts. In this approach, two control scenarios are possible:

1. Due to the congested pattern control approach, congestion dissolves and free flow recovers at the bottleneck. In this case, the congested pattern control approach can also be called “control of traffic breakdown at bottleneck.”
2. Due to the congested pattern control approach, congestion at the bottleneck is maintained at a minimum possible level.

Thus, in the congested pattern control approach, only after traffic breakdown has occurred at a network bottleneck, traffic control is automatically switched on. The main aim of the congested pattern control approach is congestion dissolution. However, if congestion dissolution is not possible to achieve, then through the use of the congested pattern control approach, the

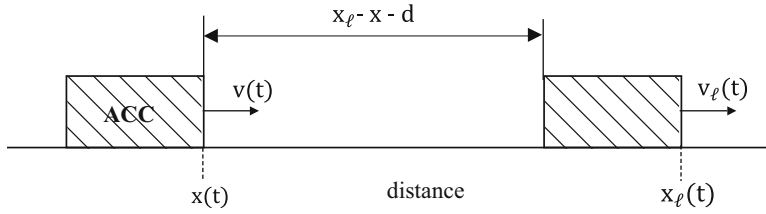
maintenance of congestion at the bottleneck should be realized at a minimum possible level. The “minimum possible level” of congestion means that traffic congestion should be localized at the bottleneck: Continuous congestion propagation upstream of the bottleneck should be prevented in the congested pattern control approach (Kerner 2004a, 2005, 2007a, b, c, 2009). In other words, rather than propagating upstream, the congested pattern should be localized at the bottleneck within a relatively small stretch of the main road.

When the congested pattern control approach is applied for on-ramp metering, it has been called ANCONA on-ramp metering “automatic on-ramp control of congested patterns” (Kerner 2004a, 2005, 2007a, b, c, 2009). ANCONA has been considered in Sects. 10.6 and 10.7 of the book (Kerner 2009) as well as the entry (Kerner 2017c) of this Encyclopedia.

Decrease in Probability of Traffic Breakdown in Networks Through Automated Driving Vehicles

A current effort of many car-developing companies is devoted to the development of automated driving vehicles. Automated driving is also called automatic driving, self-driving, or autonomous driving. It is assumed that future vehicular traffic in traffic and transportation networks is a mixed traffic flow consisting of human driving and automated driving vehicles. Automated driving vehicles should considerably enhance capacity of free flow in a traffic network. As emphasized above (section “Introduction”), capacity of free flow in the network is limited by the occurrence of traffic breakdown at a network bottleneck.

For a discussion of the effect of automated driving vehicles on traffic flow, we consider a simple case of vehicular traffic on a single-lane road with an on-ramp bottleneck. On the single-lane road, no vehicles can pass. For this reason, automated driving can be achieved through the use of an adaptive cruise control (ACC) in a vehicle: An ACC-vehicle follows a preceding vehicle automatically based on some ACC dynamics rules of motion. Depending on the



Breakdown in Traffic Networks, Fig. 8 Model of the ACC-vehicle (see, e.g. van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Chien et al. 1997; Davis 2004, 2007, 2014; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Levine and Athans 1966; Liang and

Peng 1999, 2000; McDonald and Wu 1997; Rajamani 2012; Sathiyar et al. 2013; Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001; Treiber and Helbing 2001; Treiber and Kesting 2013; Xiao and Gao 2010)

dynamic behavior of the preceding vehicle, these ACC-rules determine either automatic acceleration or automatic deceleration of the ACC-vehicle or else the maintaining of a time-independent speed. The preceding vehicle can be either a human driving vehicle or an automated driving vehicle through an ACC-system in the vehicle.

Classical Model of Automated Driving Vehicle

In simulations of mixed traffic flow on a single-lane road with an on-ramp bottleneck discussed below, there are vehicles that have no ACC-system (human driving vehicles) and ACC-vehicles (automated driving vehicles). The ACC-vehicles are randomly distributed on the road between human driving vehicles. The percentage of automated driving vehicles denoted by $\gamma^{(ACC)}$ is the same value in traffic flow upstream of the bottleneck and in the on-ramp inflow onto the main road.

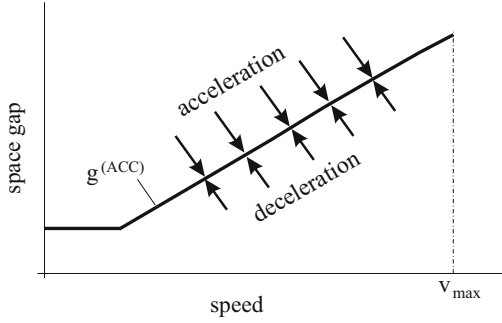
We use the classical model of ACC (see, e.g., van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Chien et al. 1997; Davis 2004, 2007, 2014; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Levine and Athans 1966; Liang and Peng 1999, 2000; McDonald and Wu 1997; Rajamani 2012; Sathiyar et al. 2013; Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001; Treiber and Helbing 2001; Treiber and Kesting 2013; Xiao and Gao 2010):

$$a^{(ACC)}(t) = K_1 \left(g(t) - v(t)\tau_d^{(ACC)} \right) + K_2 (v_\ell(t) - v(t)), \quad (9)$$

where K_1 and K_2 are coefficients of ACC adaptation (In real ACC, the coefficients of ACC

adaptation can be functions of the speed and the space gap. Moreover, the coefficients of ACC adaptation can depend on a driving situation. However, the objective of this section is a study of *qualitative* features of the effect of ACC-vehicles in mixed traffic flow on the probability of traffic breakdown at network bottlenecks. To reach this goal, we can assume that the coefficients of ACC adaptation do not depend on the speed and the space gap in mixed traffic flow.), $v(t)$ is the speed of the ACC-vehicle, $v_\ell(t)$ is the speed of the preceding vehicle (Fig. 8), $\tau_d^{(ACC)}$ is a desired net time gap (desired time headway) of the ACC-vehicle to the preceding vehicle, and $a^{(ACC)}(t)$ is acceleration (deceleration) of the ACC-vehicle that is determined by the space gap to a preceding vehicle $g(t) = x_\ell(t) - x(t) - d$ and the relative speed $\Delta v(t) = v_\ell(t) - v(t)$ measured by the ACC-vehicle as well as by a desired space gap $g^{(ACC)} = v(t)\tau_d^{(ACC)}$.

In Eq. (9), coefficients of ACC adaptation K_1 and K_2 have the following physical meaning (see, e.g., works van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Chien et al. 1997; Davis 2004, 2007, 2014; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Levine and Athans 1966; Liang and Peng 1999, 2000; McDonald and Wu 1997; Rajamani 2012; Sathiyar et al. 2013; Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001; Treiber and Helbing 2001; Treiber and Kesting 2013; Xiao and Gao 2010 and references there): The coefficients K_1 and K_2 describe the dynamic adaptation of the ACC-vehicle when either the space gap g is different from the desired one $g^{(ACC)} = v\tau_d^{(ACC)}$ (Fig. 9), i.e.,



Breakdown in Traffic Networks, Fig. 9 Operating points of the classical model of ACC: Qualitative speed dependence of desired space gap $g^{(ACC)} = v\tau_d^{(ACC)}$ (see, e.g. van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Chien et al. 1997; Davis 2004, 2007, 2014; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Levine and Athans 1966; Liang and Peng 1999, 2000; McDonald and Wu 1997; Rajamani 2012; Sathiyar et al. 2013; Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001; Treiber and Helbing 2001; Treiber and Kesting 2013; Xiao and Gao 2010). When the speed of the ACC-vehicle is smaller than a speed associated with a given minimum space gap between vehicles, then the desired space gap does not depend on the speed as shown in figure for small enough speeds; we do not consider this speed range in the main text

$$g \neq v\tau_d^{(ACC)} \quad (10)$$

or the vehicle speed is different from the speed of the preceding vehicle, i.e.,

$$v \neq v_\ell. \quad (11)$$

In contrast, if the time headway of the ACC-vehicle $\tau^{(net)}$ is equal to the desired time headway

$$\tau^{(net)} = \tau_d^{(ACC)} \quad (12)$$

and the relative speed $\Delta v = v_\ell - v$ of the ACC-vehicle v to the speed of the preceding vehicle v_ℓ is zero.

$$\Delta v = 0, \quad (13)$$

then from (9) we obtain that.

$$a^{(ACC)} = 0. \quad (14)$$

This means that the ACC-vehicle moves with a time-independent speed.

Conditions (12) and (13) determine the so-called operating points of the ACC-vehicle (see, e.g., van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Liang and Peng 1999, 2000; Rajamani 2012; Sathiyar et al. 2013; Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001). An operating point of the ACC-vehicle is a hypothetical steady state of the ACC-vehicle that follows the preceding vehicle moving at a time-independent speed. In the operating point, both the acceleration (deceleration) of the ACC-vehicle and the relative speed of the ACC-vehicle to the speed of the preceding vehicle are equal to zero. Accordingly (12), in the classical model of the ACC-vehicle at each given speed of the ACC-vehicle that is larger than zero, the operating point of the ACC-vehicle corresponds to a time headway that is equal to the desired time headway $\tau_d^{(ACC)}$. For each of the operating points of the ACC-vehicle, formula (12) is equivalent to formula.

$$g^{(ACC)} = v\tau_d^{(ACC)}. \quad (15)$$

Using formula (15), the operating points of the classical ACC (12) and (13) can be presented by a curve in the gap–speed flow plane (Fig. 9). For each given speed $v > 0$ of the ACC-vehicle, there is only one operating point of the classical ACC that is determined by formulas (12) and (13) (see, e.g., works van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Liang and Peng 1999, 2000; Rajamani 2012; Sathiyar et al. 2013; Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001 and references there).

- For each given speed of the ACC-vehicle, there is only one operating point for the classical ACC.

The physics of the dynamic equation for the ACC-vehicle (9) is as follows (see, e.g., works van Arem et al. 2006; Bose and Ioannou 2003; Cheng 2011; Ioannou and Chien 1993; ISO 15622 2010; Kukuchi et al. 2003; Liang and Peng 1999, 2000; Rajamani 2012; Sathiyar et al. 2013;

Shladover 1995; Swaroop and Hedrick 1996; Swaroop et al. 2001 and references there). It can be seen that in Eq. (9) the current time headway $\tau^{(\text{net})} = g/v$ is compared with the desired time headway $\tau_d^{(\text{ACC})}$. If

$$\tau^{(\text{net})} > \tau_d^{(\text{ACC})}, \quad (16)$$

then the ACC-vehicle automatically accelerates to reduce the time headway $\tau^{(\text{net})}$ to the desired value $\tau_d^{(\text{ACC})}$ (labeled by “acceleration” in Fig. 9). If

$$\tau^{(\text{net})} < \tau_d^{(\text{ACC})}, \quad (17)$$

then the ACC-vehicle decelerates automatically to increase the time headway (labeled by “deceleration” in Fig. 9). Moreover, the acceleration and deceleration of the ACC-vehicle depend on the current difference between the speed of the ACC-vehicle and the preceding vehicle. If the preceding vehicle has a higher speed than the ACC-vehicle, i.e., when,

$$v_\ell > v, \quad (18)$$

the ACC-vehicle accelerates. Otherwise, if,

$$v_\ell < v, \quad (19)$$

the ACC-vehicle decelerates.

An important characteristic of the ACC-vehicles is a stability of a platoon of the ACC-vehicles called *string stability*. Liang and Peng (1999) have found that for a string stability of the ACC-vehicles coefficients of ACC adaptation K_1 and K_2 in (9) and the desired time headway of the ACC-vehicles, $\tau_d^{(\text{ACC})}$ should satisfy condition (Liang and Peng 1999)

$$K_2 > \frac{2 - K_1 \left(\tau_d^{(\text{ACC})} \right)^2}{2\tau_d^{(\text{ACC})}}. \quad (20)$$

Probability of Traffic Breakdown at Network Bottlenecks in Mixed Traffic Flow

In this section, following Kerner (2017b) we consider the answer on the following question:

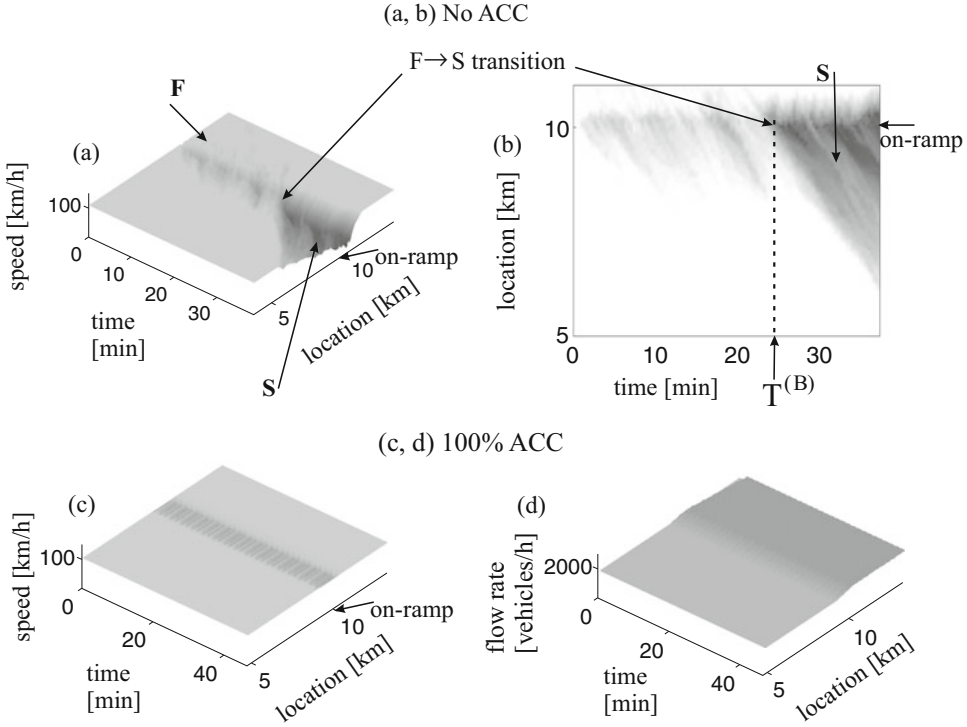
- Can automated driving vehicles in mixed traffic flow reduce the probability of traffic breakdown at network bottlenecks?

The classical theories failed in the understanding of stochastic highway capacity (section “[Classical Understanding of Stochastic Highway Capacity](#)”). For this reason, as explained in Chap. 4 of the book (Kerner 2017b), it is not possible to perform a reliable study of the effect of automated driving vehicles in mixed traffic flow on the probability of traffic breakdown at network bottlenecks with the use of the classical traffic flow models. Therefore, for the analysis of the effect of automated driving vehicles on traffic flow, we use simulations of the Kerner-Klenov microscopic stochastic traffic flow model (see Appendix A of the book Kerner 2017b) that is in the framework of the three-phase theory. The model can explain the empirical nucleation nature of traffic breakdown.

We consider traffic flow of manual driving vehicles on single-lane road with an on-ramp bottleneck at the flow rates at the bottleneck $q_{\text{on}} = 320$ vehicles/h and $q_{\text{in}} = 2000$ vehicles/h at which after a random time delay $T^{(\text{B})}$ traffic breakdown occurs (Fig. 10a, b). At the same set of the flow rates $q_{\text{on}}, q_{\text{in}}$ as those in Fig. 10a, b, in traffic flow consisting of 100% of ACC-vehicles that satisfy condition (20) for string stability, no traffic breakdown can occur (Fig. 10c, d).

We have found that as long as the percentage of ACC-vehicles is appreciably smaller than $\gamma^{(\text{ACC})} = 10\%$, no considerable change in the probabilistic features of traffic breakdown at the bottleneck is observed. Even when the percentage of ACC-vehicles increases to $\gamma^{(\text{ACC})} = 10\%$, features of traffic breakdown remain almost the same as those in traffic flow of manual driving vehicles (Fig. 10a, b), and only a relatively small decrease in the probability of the breakdown is observed (compare curves labeled by “10% ACC” and “No ACC” in Fig. 11a).

The physics of the decrease of the probability of traffic breakdown in mixed traffic flow is as follows. In traffic flow consisting of human driving vehicles only, traffic breakdown ($F \rightarrow S$ transition) occurs (Fig. 10a, b), when a local speed disturbance that causes a large enough local speed



Breakdown in Traffic Networks, Fig. 10 Traffic breakdown in traffic flow of manual driving vehicles on single-lane road with an on-ramp bottleneck (a, b) and stable free flow of 100% of ACC-vehicles (c, d): Vehicle speed (a, c) and the flow rate (averaging over 20 vehicles) (d) in space and time. (b) The same speed data as those in (a) but presented by regions with variable shades of gray (shades

of gray vary from white to black when the speed decreases from 105 km/h (white) to zero (black)). Parameters of ACC-vehicles: $\tau_d^{(ACC)} = 1.1$ s and coefficients $(K_1, K_2) = (0.14 \text{ s}^{-2}, 0.9 \text{ s}^{-1})$ satisfying (20). Flow rates in all figures are $q_{on} = 320$ vehicles/h, $q_{in} = 2000$ vehicles/h. F – free flow, S – synchronized flow. On-ramp location $x_{on} = 10$ km (Taken from Kerner 2017b)

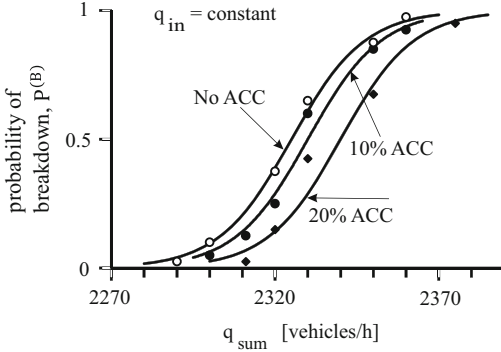
decrease (nucleus for the breakdown) appears randomly in metastable free flow at the bottleneck. Free flow that consists of 100% automated driving vehicles is stable (Fig. 10c, d). For this reason, we can assume that a long enough platoon of ACC-vehicles that propagates through a local disturbance at the bottleneck can prevent the growth of the local disturbance. The larger the percentage $\gamma^{(ACC)}$ of automated driving vehicles in mixed traffic flow, the larger the probability of the appearance of the long platoon of ACC-vehicles, therefore, the larger the probability of the prevention of the disturbance growth and the smaller the breakdown probability $P^{(B)}$ are.

These qualitative explanations are confirmed by numerical simulations of the flow rate dependence of the probability of traffic breakdown $P^{(B)}(q_{sum})$

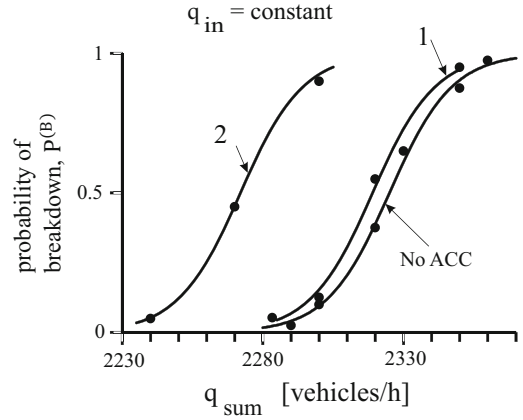
presented in Fig. 11. Indeed, we have found that the flow rate dependence of the probability of traffic breakdown for mixed traffic flow with 20% automated driving vehicles (right curve $P^{(B)}(q_{sum})$ labeled by “20% ACC” in Fig. 11) is shifted to larger flow rates in comparison with the function $P^{(B)}(q_{sum})$ for traffic flow consisting of human driving vehicles only (left curve labeled by “No ACC” in Fig. 11).

Deterioration of Performance of Traffic System Through Automated Driving Vehicles

Rather than the enhancement of traffic flow characteristics (section “Decrease in Probability of



Breakdown in Traffic Networks, Fig. 11 Probabilities of traffic breakdown $P^{(B)}(q_{\text{sum}})$ in traffic flows without ACC-vehicles (left curve labeled by “No ACC”) as well as with 10% and 20% of ACC-vehicles (right curves labeled by “10% ACC” and “20% ACC”) as functions of the flow rate downstream of the bottleneck q_{sum} . The observation time of traffic flow is equal to $T^{(B)} = 30$ min. The flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ is varied through the change in the on-ramp inflow rate q_{on} at constant $q_{\text{in}} = 2000$ vehicles/h. Other model parameters are the same as those in Fig. 10 (Taken from Kerner 2017b)



Breakdown in Traffic Networks, Fig. 12 Increase in the probability of traffic breakdown $P^{(B)}$ in mixed free traffic flow through ACC-vehicles with $\gamma_d^{(\text{ACC})} = 1.1$ s: The flow rate dependence of the breakdown probability $P^{(B)}(q_{\text{sum}})$ labeled by “No ACC” is taken from Fig. 11 for free flow of manual driving vehicles without ACC-vehicles. Curves $P^{(B)}(q_{\text{sum}})$ labeled by numbers 1 and 2 are related to mixed free flows with $\gamma^{(\text{ACC})} = 5\%$ of ACC-vehicles with different sets of coefficients $(K_1, K_2) = (0.2 \text{ s}^{-2}, 0.82 \text{ s}^{-1})$ for curve 1, $(0.7 \text{ s}^{-2}, 0.55 \text{ s}^{-1})$ for curve 2, which satisfy (20) (Taken from Kerner 2017b)

Traffic Breakdown in Networks Through Automated Driving Vehicles”), automated driving vehicles can also result in the deterioration of the performance of the traffic system (Kerner 2017b): Automated driving vehicles can increase the probability of traffic breakdown at network bottlenecks considerably.

It turns out that already a single automated driving vehicle can initiate traffic breakdown at a network bottleneck. To show the deterioration of the performance of the traffic system through automated driving vehicles, we continue to consider automated driving vehicles that satisfy condition (20) for string stability. Moreover, we consider the flow rates and parameters of automated driving vehicles at which no traffic breakdown can occur at the bottleneck when traffic flow consists of 100% of automated driving vehicles. The ACC-vehicles exhibit the same short desired time headway $\tau_d^{(\text{ACC})} = 1.1$ as used above in section “Decrease in Probability of Traffic Breakdown in Networks Through Automated Driving Vehicles.” Nevertheless, we will find that even such ACC-vehicles while moving in mixed traffic flow can lead to a considerable increase in the probability of traffic breakdown at road bottlenecks (Fig. 12).

Indeed, we have found that even a relatively small percentage $\gamma^{(\text{ACC})} = 5\%$ of the ACC-vehicles in mixed traffic flow can increase considerably the probability of traffic breakdown (Fig. 12, curves 1 and 2). The importance of this result is as follows: In section “Decrease in Probability of Traffic Breakdown in Networks Through Automated Driving Vehicles,” we have mentioned that the positive effect of the ACC-vehicles on traffic flow, in particular, the decrease in the probability of traffic breakdown, is considerable only at large percentages of ACC-vehicles $\gamma^{(\text{ACC})} \approx 20\%$. In the next future, we could expect only much smaller percentages of automated driving vehicles in mixed traffic flow, like $\gamma^{(\text{ACC})} = 1 - 5\%$. Therefore, the deterioration of the performance of mixed traffic flow shown in Fig. 12 (curves 1 and 2) can be a subject of the development of automated driving vehicles in car-development companies already during next years.

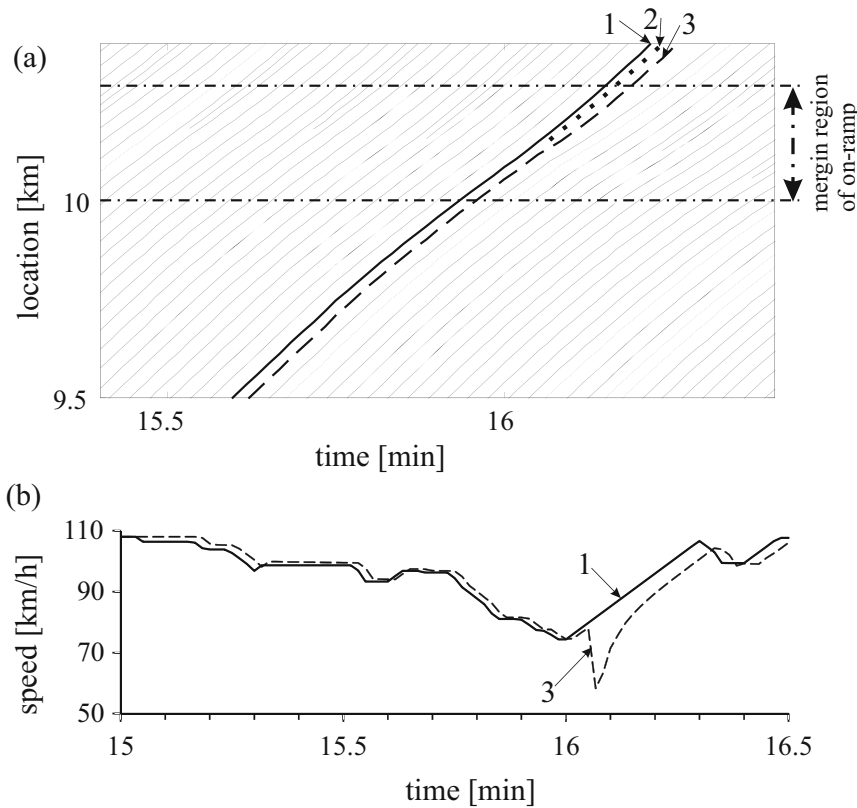
The physical reason for the deterioration of the performance of the traffic system through automated driving vehicles found out in simulations (Kerner 2016a) and presented in Chap. 6 of the

book (Kerner 2017b) can be explained as follows. When a slow moving manual driving vehicle (dotted vehicle trajectory 2 in Fig. 13) merges from the on-ramp onto the main road at the bottleneck, there is no effect of this vehicle merging on the motion of downstream vehicle 1 (solid vehicle trajectory 1 in Fig. 13). In contrast, an ACC-vehicle (dashed vehicle trajectory 3 in Fig. 13) that moves on the main road upstream of vehicle 2 decelerates strongly while adapting the speed to the speed of vehicle 2. It turns out that this strong deceleration of the automated driving vehicle (ACC-vehicle) causes a large speed disturbance at the bottleneck (dashed vehicle trajectory 3 in Fig. 13). On average, the speed disturbance at the bottleneck is considerably larger than the disturbance occurring through

manual driving vehicles. Simulations of the effect of automated driving vehicles on mixed traffic flow presented in Chap. 6 of the book Kerner (2017b) lead to the following result:

- The deterioration of the performance of the traffic system through automated driving vehicles can occur within broad ranges of the percentage of ACC-vehicles and the set of coefficients (K_1 , K_2), which satisfy condition (20) of string stability of ACC-vehicles.

It should be noted that the desired time headway of the ACC-vehicle $\tau_d^{(ACC)} = 1.1$ s used in simulations (Figs. 10, 11, 12, and 13) is a relatively short one. In real mixed traffic, longer values of the desired time headway of the ACC-

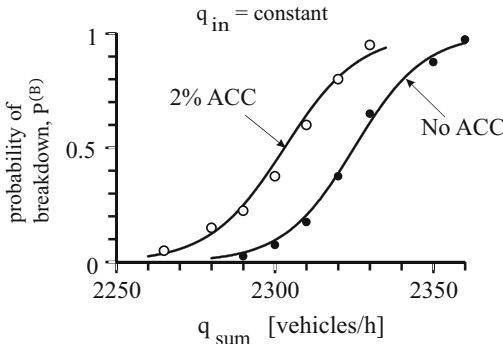


Breakdown in Traffic Networks, Fig. 13 Simulations of dynamics of permanent local speed disturbance at on-ramp bottleneck on single-lane road for mixed traffic flow with $\gamma^{(ACC)} = 5\%$ of ACC-vehicles with $\gamma_d^{(ACC)} = 1.1$ s and coefficients $(K_1, K_2) = (0.5 \text{ s}^{-2}, 0.65 \text{ s}^{-1})$: (a)

Fragment of vehicle trajectories in space and time. (b) Microscopic vehicle speeds along trajectories as time functions labeled by the same numbers as those in (a). $q_{on} = 320$ vehicles/h, $q_{in} = 2000$ vehicles/h (Taken from Kerner 2017b)

vehicle $\tau_d^{(ACC)} = 1.2 - 1.5$ s can be expected. Simulations show if ACC-vehicles with $\tau_d^{(ACC)} = 1.1$ s lead to the increase in the breakdown probability, then at the same chosen set of coefficients (K_1, K_2) of the ACC-vehicles, an increase in $\tau_d^{(ACC)}$ causes an increase in the breakdown probability: The longer the desired time headway of the ACC-vehicle $\tau_d^{(ACC)}$, the larger the breakdown probability. At a long enough value $\tau_d^{(ACC)}$, already a single automated driving vehicle causes a large speed disturbance at a network bottleneck that initiates traffic breakdown at the bottleneck. In this case, already at a very small percentage of automated driving vehicles (about 1–2%) in mixed traffic flow, the considerable increase in the probability of traffic breakdown at the bottleneck has been found.

An example is shown in Figs. 14 and 15. We have found that at a slightly longer desired time headway of the ACC-vehicle $\tau_d^{(ACC)} = 1.3$ s in comparison with the value $\tau_d^{(ACC)} = 1.1$ s used in Fig. 13, already 2% of automated driving vehicles in mixed traffic flow increases the probability of traffic breakdown considerably in comparison with traffic flow without automated driving vehicle (Fig. 14). This is because the longer the desired time headway of the ACC-vehicle $\tau_d^{(ACC)}$, the

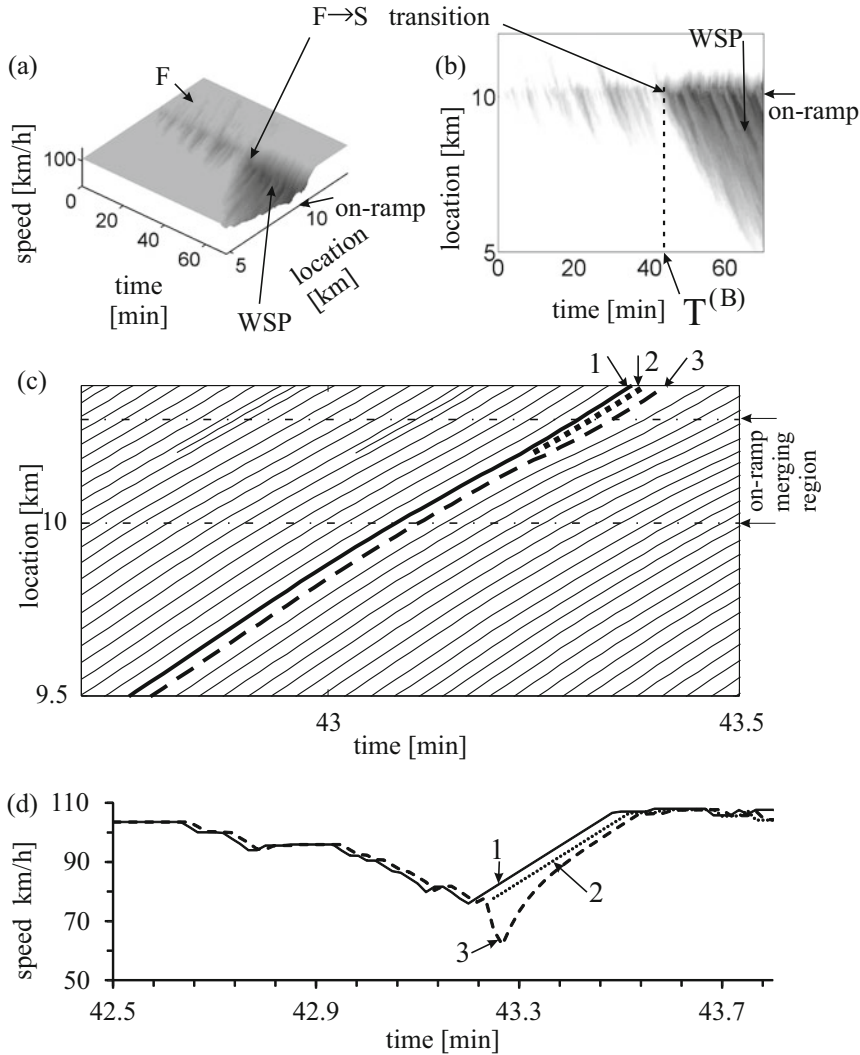


Breakdown in Traffic Networks, Fig. 14 Increase in the probability of traffic breakdown $P^{(B)}$ in mixed free traffic flow through ACC-vehicles with $\gamma^{(ACC)} = 2\%$ of ACC-vehicles with $\tau_d^{(ACC)} = 1.3$ s and coefficients (K_1, K_2) = (0.3 s⁻², 0.6 s⁻¹) (curve labeled by “2% ACC”). The flow rate dependence of the breakdown probability $P^{(B)}(q_{sum})$ labeled by “No ACC” is taken from Fig. 11 for free flow of manual driving vehicles without ACC-vehicles

stronger the deceleration of the ACC-vehicle as the reaction on the merging of a slower moving vehicle from the on-ramp. For this reason, already a single automated driving vehicle can lead to traffic breakdown as shown in Fig. 15. As in Fig. 13, in this case the automated driving vehicle (dashed trajectory 3 in Fig. 15c, d) decelerates strongly within the merging region of the on-ramp due to the merging of a slower moving vehicle 2 from the on-ramp (dotted trajectory 2 in Fig. 15c, d). This deceleration of the automated driving vehicle causes the occurrence of a nucleus for traffic breakdown (F → S transition) with the resulting formation of a widening synchronized flow pattern (WSP) at the bottleneck (Fig. 15a, b).

Thus, these simulations (Fig. 15) prove that a strong deceleration of a single automated driving vehicle in a neighborhood of a network bottleneck can be the source for the nucleus occurrence leading to traffic breakdown at the bottleneck. We can make the following conclusions:

- Depending on the parameters of automated driving vehicles and on the percentage of the automated driving vehicles in mixed traffic flow consisting of a random distribution of automated driving and manual driving vehicles, the automated driving vehicles can either decrease or increase the probability of traffic breakdown in mixed traffic flow.
- The effect of the deterioration of the performance of the traffic system (increase in the probability of traffic breakdown) through automated driving vehicles occurs even if no breakdown can occur in traffic flow consisting of 100% automated driving vehicles and the automated driving vehicles satisfy condition (20) of string stability.
- For automated driving vehicles that can decrease the breakdown probability in mixed traffic flow, the following result is valid: To decrease noticeably the probability of traffic breakdown at a network bottleneck, a large percentage of automated driving vehicles (about and more than 20%) is required.
- In contrast, for automated driving vehicles that can increase the breakdown probability in mixed traffic flow, the following result has



Breakdown in Traffic Networks, Fig. 15 Simulations of dynamics of permanent local speed disturbance at on-ramp bottleneck on single-lane road for mixed traffic flow with $\gamma^{(ACC)} = 2\%$ of ACC-vehicles with $\tau_d^{(ACC)} = 1.3$ s and coefficients $(K_1, K_2) = (0.3 \text{ s}^{-2}, 0.6 \text{ s}^{-1})$: (a) Vehicle speed in space and time. (b) The same speed data as those in (a) but presented by regions with variable shades

of gray (shades of gray vary from white to black when the speed decreases from 105 km/h (white) to zero (black)). (c) Fragment of vehicle trajectories in space and time. (d) Microscopic vehicle speeds along trajectories as time functions labeled by the same numbers as those in (c). $q_{on} = 280$ vehicles/h, $q_{in} = 2000$ vehicles/h

been found: The deterioration of the performance of the traffic system through automated driving vehicles can be realized at a small percentage of automated driving vehicles (about 1–5%) in mixed traffic flow. This is because already a single automated driving vehicle can initiate a nucleus required for traffic breakdown at the bottleneck.

Automated Driving Based on Three-Phase Theory

The deterioration of the performance of the traffic system through the ACC-vehicles discussed in section “[Deterioration of Performance of Traffic System Through Automated Driving Vehicles](#)” could be avoided through the use of automated

driving vehicles, which learn from behaviors of drivers in real traffic as incorporated in hypotheses of the three-phase theory.

To understand this statement, we should firstly recall that for each given speed $v > 0$ of a classical ACC-vehicle, there is *only one* operating point of the ACC that is determined by formula $\tau^{(\text{net})} = \tau_d^{(\text{ACC})}$ (12): When both the preceding vehicle and the ACC-vehicle move at the same time-independent speed $v = v_\ell = \text{const} > 0$, then the time headway $\tau^{(\text{net})}$ of the classical ACC-vehicle is steered to a given fixed time headway $\tau_d^{(\text{ACC})}$ chosen by a driver that is also called a desired time headway. The operating points of the classical ACC (12) and (13) related to different values of the speed of the classical ACC can be presented by a curve in the gap-speed flow plane (Fig. 9).

Automated Driving Without Fixed Desired Time Headway to Preceding Vehicle

In contrast with this behavior of the classical ACC-vehicle, from hypothesis of the three-phase theory about 2D-steady states of synchronized flow (Kerner 1998a, b, 1999a, b), it follows that drivers do not try to maintain some fixed time headway to the preceding vehicle while moving at a nearly time-independent speed. Therefore, one of the main features of real drivers that automated driving vehicles should learn from behaviors of drivers in real traffic is the driver behavior incorporated in the hypothesis about 2D-states of synchronized flow of the three-phase theory.

In contrast with the classical model of the ACC-vehicle (9), as abovementioned, for the ACC-vehicle based on the three-phase theory, there is *no* fixed desired time headway of the ACC-vehicle to the preceding vehicle. We call automated driving vehicles in which there is no fixed desired time headway to the preceding vehicle as “automated driving based on the three-phase theory.”

To understand the difference between the classical model of ACC (9) and the model of automated driving based on the three-phase theory (Kerner 2004b, 2007d, e, f, 2008), we assume that the space gap of the ACC-vehicle g to the preceding vehicle satisfies conditions.

$$g_{\text{safe}} \leq g \leq G, \quad (21)$$

where G and g_{safe} are the synchronization and safe space gaps, respectively (Fig. 16). We assume that the speed of the ACC-vehicle is higher than zero ($v > 0$). Then, formula (21) is equivalent to conditions:

$$\tau_{\text{safe}} \leq \tau^{(\text{net})} \leq \tau_G, \quad (22)$$

where $\tau_{\text{safe}} = g_{\text{safe}}/v$, $\tau_G = G/v$.

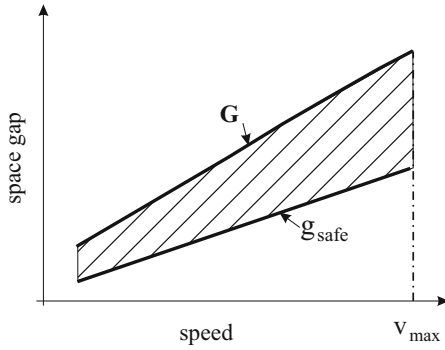
We consider possible operating points of the ACC based on the three-phase theory. An operating point of the ACC is related to the case, when the speed difference between the speed of the preceding vehicle and the speed of the ACC-vehicle $\Delta v = v_\ell - v$ is equal to zero (13) and the preceding vehicle moves at a time-independent speed v_ℓ (where we assume that $v_\ell > 0$). Indeed, in this case in accordance with the definition of an operating point of the ACC-vehicle (section “[Classical Model of Automated Driving Vehicle](#)”), the ACC-vehicle moves at the time-independent speed that is equal to the speed of the preceding vehicle: $v = v_\ell$. From conditions (22), we get that for the ACC based on the three-phase theory moving at a time-independent speed $v = v_\ell > 0$, an operating point of the ACC should satisfy conditions.

$$\tau_{\text{safe}}(v) \leq \tau^{(\text{net})} \leq \tau_G(v), \quad (23)$$

where we have taken into account that in a general case, the safe time headway τ_{safe} and/or the synchronization time headway τ_G can be speed functions. Additionally, in (23) for each given speed v , the safe time headway $\tau_{\text{safe}}(v)$ and the synchronization time headway $\tau_G(v)$ are some constant values that satisfy condition.

$$\tau_G(v) > \tau_{\text{safe}}(v). \quad (24)$$

Conditions (23) for the automated driving vehicles based on the three-phase theory are the same ones as conditions for manual driving vehicles following from the hypothesis of the three-phase theory about 2D-steady states of synchronized flow (Kerner 1998a, b, 1999a, b) (see the



Breakdown in Traffic Networks, Fig. 16 Qualitative explanation of the infinite number of the operating points for the ACC based on the three-phase theory (Kerner 2004b, 2007d, e, f, 2008): A part of 2D-region for operating points of the ACC in the space gap–speed plane (dashed region) that satisfies conditions (21). In contrast, in the classical model of ACC (section “Classical Model of Automated Driving Vehicle”) at the same time-independent speed, there is only one operating point of the ACC (see Fig. 9)

books (Kerner 2004a, 2009, 2017b) and entry (Kerner 2017c) of this Encyclopedia).

We can see that for an automated driving vehicle based on the three-phase theory (Fig. 16) (Kerner 2004b, 2007d, e, f), there should be at least one speed range within which for each given speed that is larger than zero, there is an infinite number of different time headway $\tau^{(\text{net})}$ that satisfy conditions (23).

- We define an automated driving vehicle based on the three-phase theory (Kerner 2004b, d, e, f) as the automated driving vehicle for which there is a speed range within which for each given vehicle speed there is an infinite number of the operating points.

However, the question arises:

- What can be the dynamic behavior of automated driving vehicle based on the three-phase theory?

We consider a usual case at which the speed of the ACC-vehicle $v(t)$ and the speed of the preceding vehicle $v_\ell(t)$ are time functions and the speed of the ACC-vehicle is not equal to the speed of the preceding vehicle: $v(t) \neq v_\ell(t)$. When conditions

(21) (or (22)) are satisfied, the space gap of the ACC-vehicle to the preceding vehicle is within a 2D region in the space gap–speed plane (dashed region in Fig. 17). Then, rather than the classical formula (9), the acceleration (deceleration) of automated driving vehicles based on the three-phase theory is given by formula (Kerner 2004b, 2007d, e, f).

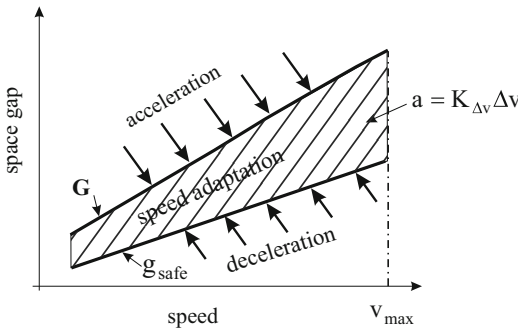
$$a(t) = K_{\Delta v} \Delta v(t), \quad (25)$$

where $\Delta v(t) = v_\ell(t) - v(t)$, $v(t)$ is the speed of the ACC-vehicle at time instant t , $v_\ell(t)$ is the speed of the preceding vehicle at time instant t , and $K_{\Delta v}$ is a dynamic coefficient of the ACC-vehicle ($K_{\Delta v} > 0$).

Formula (25) means that the ACC-vehicle adapts its speed $v(t)$ to the speed of the preceding vehicle $v_\ell(t)$ without caring, what the precise space gap (time headway) to the preceding vehicle is. Thus, the dynamics of the automated driving vehicle based on the three-phase theory should not necessarily depend on the time headway to the preceding vehicle $\tau^{(\text{net})}(t)$ as long as conditions (22) are satisfied. We can also formulate this conclusion as follows:

- An automated driving vehicle based on the three-phase theory (Kerner 2004b, 2007d, e, f) is the automated driving vehicle in which there is no fixed desired time headway to the preceding vehicle.

Outside the 2D-region in the space gap–speed plane (Fig. 17) formula (25) is not applied (Kerner 2007e, f, 2008): At $g > G$, i.e., $\tau^{(\text{net})} > \tau_G$, the ACC-vehicle accelerates (labeled by “acceleration” in Fig. 17), whereas at $g < g_{\text{safe}}$, i.e., $\tau^{(\text{net})} < \tau_{\text{safe}}$, the ACC-vehicle decelerates (labeled by “deceleration” in Fig. 17) (In real ACC based on the three-phase theory, the synchronization time headway τ_G , the safe time headway τ_{safe} , and the coefficient of ACC adaptation $K_{\Delta v}$ in (25) can be complex functions of the speed of the vehicle and the speed of the preceding vehicle as well as of the current space gap (current time headway). Moreover, these functions can depend on a driving situation. Some examples of these functions can be found in Kerner 2007e, f, 2008.).



Breakdown in Traffic Networks, Fig. 17 Explanation of ACC based on three-phase theory (Kerner 2004b, 2007d, e, f, 2008): A part of 2D-region for operating points of the ACC in the space gap–speed plane (dashed region taken from Fig. 16) within which an ACC-vehicle moves in accordance with Eq. (25)

- For the automated driving vehicle based on the three-phase theory, there is no fixed time headway to which the automated driving vehicle should necessarily be steered (Kerner 2004b, 2007d, e, f).

Thus, in some traffic situations, acceleration (deceleration) of the ACC-vehicle based on the three-phase theory does not depend on the space gap, i.e., on the time headway to the preceding vehicle at all.

In 2007–2008, an ACC based on the three-phase theory, in which formula (25) under conditions (21) and (22) was used, was implemented in real ACC-vehicles and proven in test drives by the Daimler Company. It was found that the ACC-vehicle based on the three-phase theory exhibits a more comfortable ACC behavior as well as it leads to the reduction of fuel consumption and CO₂ emissions. To the knowledge of the author, ACC-vehicles based on the three-phase theory (Kerner 2004b, 2007d, e, f, 2008) are currently on the market. The effect of automated driving based on the three-phase theory on mixed traffic flow has been revealed in (Kerner 2017d).

Possible advantages of the ACC-system based on the three-phase theory in comparison with the classical ACC-system (9) are as follows:

- The prevention of the deterioration of the performance of the traffic system through automated driving vehicles.

- The reducing of a conflict between the dynamic and comfortable ACC behavior. In particular, at the same other ACC dynamic behaviors, a much comfortable ACC-vehicle motion is possible.
- The reduction of fuel consumption and CO₂ emissions while moving in congested traffic.
- Through the use of condition (25), automated driving vehicles can alleviate the problem of large disturbances occurring during vehicle merging at a bottleneck. For this reason, such automated driving vehicles can prevent traffic breakdown at the bottleneck.

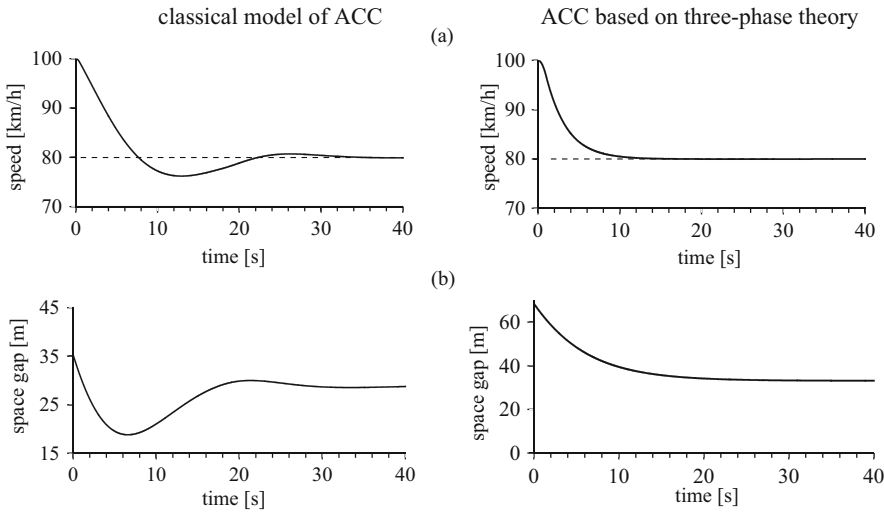
A more detailed analysis of autonomous driving vehicles in mixed traffic flow can be found in the article ► [“Autonomous Driving in the Framework of Three-Phase Traffic Theory”](#) of this volume (pages 343–385).

Approaching an Operating Point of Automated Driving Vehicle

To illustrate one of the differences in the dynamic behavior of an automated driving vehicle based on the three-phase theory in comparison with the dynamic behavior of the classical model (9) of automated driving vehicles, we consider a simple scenario in which an automated driving vehicle approaches a slower moving preceding vehicle that cannot be overtaken. Simulations of a simple case in which the preceding vehicle moves at a time-independent speed are presented in Fig. 18.

As well known, by approaching the preceding vehicle automated driving vehicle whose dynamics is determined by classical model (9) can exhibit oscillating behavior (Fig. 18, left panel): Before the automated driving vehicle reaches the operating point, there are decaying oscillations both in the speed and space gap in the vicinity of, respectively, the speed and space gap related to the operating point. The amplitude and duration of the oscillations depend on coefficients K_1 and K_2 in (9).

Contrarily to automated driving vehicles whose dynamics is determined by classical model of ACC (9) (Fig. 18, left panel), the dynamic of the approaching of an operating point of an automated driving vehicle based on the three-phase theory exhibits *no* oscillations in the speed and space gap (Fig. 18, right panel).



Breakdown in Traffic Networks, Fig. 18 Simulations of approaching an operating point of automated driving vehicle whose dynamics is determined by classical model (9) with $\tau_d^{(ACC)} = 1.3$ s and coefficients $(K_1, K_2) = (0.06 \text{ s}^{-2}, 0.2 \text{ s}^{-1})$ (left panel) and automated driving vehicle

based on three-phase theory (right panel): (a, b) Time dependencies of the speed (a) and space gap (b) for automated driving vehicle approaching the preceding vehicle moving at time-independent speed $v = 80$ km/h

This aperiodic dynamic approaching of one of the infinity of the operating points of the automated driving vehicle based on the three-phase theory is explained by the speed adaptation effect of the three-phase theory: At each vehicle speed, the automated driving vehicle based on the three-phase theory can make an arbitrary choice in the time headway that satisfies conditions (23). This means that the automated driving vehicle based on the three-phase theory accepts different values of time headway gaps at different times and does not control a fixed time headway to the preceding vehicle.

Random Time-Delayed Breakdown in City Traffic

In accordance with the classical theory of traffic at the signal (see, e.g., Gartner (1983), Gartner and Stamatiadis (2002), Grafton and Newell (1967), Hunt et al. (1981), Little et al. (1982), Morgan and Little (1964), Webster (1958) and references in reviews (Dion et al. 2004; Gartner and Stamatiadis 2009; Newell 1982), traffic breakdown (transition from undersaturated traffic to oversaturated traffic) at the signal should occur

without any delay when the average arrival flow rate at the signal exceeds a classical signal capacity.

In contrast with the classical model of traffic breakdown at the signal, recently we have found (Kerner 2011b, 2013b, 2014) that at the same value of the average arrival flow rate at the signal that exceeds the classical signal capacity, there can be two states of traffic at the signal: (i) A metastable state of undersaturated traffic and (ii) a state of oversaturated traffic. Due to the existence of the metastable undersaturated traffic, the transition from under- to oversaturated traffic at the signal is a probabilistic effect that exhibits the nucleation nature. For this reason, there can be a random time delay before the transition from undersaturated traffic to oversaturated traffic occurs at the signal (Kerner 2011b, 2013b, 2014). The time delay of the transition from under- to oversaturated traffic at the signal can be equal to several signal cycles.

A brief consideration of features of the time-delayed traffic breakdown at the signal is the objective of this section. However, firstly in section “Classical Capacity of Traffic Signal,” we consider main conclusions of the classical theory of traffic breakdown at the signal.

Classical Capacity of Traffic Signal

Results of this section “[Classical Capacity of Traffic Signal](#)” are based on a brief consideration of the classical Webster’s model of traffic at the signal (Webster 1958) as well as the further development of the Webster’s model made by several generations of traffic researchers (see, e.g., Dion et al. 2004; Gartner 1983; Gartner and Stamatiadis 2002, 2009; Grafton and Newell 1967; Hunt et al. 1981; Little et al. 1982; Morgan and Little 1964; Newell 1982).

Traffic breakdown at the signal is the transition from under- to oversaturated traffic. In undersaturated traffic at the signal, all vehicles that have come to a stop at the signal during the red signal phase can pass the signal location during the green signal phase. In other words, the queue of vehicles dissolves fully during the green signal phase (Fig. 19a).

In contrast, in oversaturated traffic at the signal, the queue of vehicles built during the red signal phase does not dissolve fully during the green signal phase (Fig. 19c). As a result, at least one of the vehicles must wait at the signal longer than one signal cycle to pass the signal. In other words, some of the vehicles that have come to a stop at the signal during the red signal phase cannot pass the signal location during the green signal phase.

In the classical theories of city traffic, it is assumed (Dion et al. 2004; Gartner 1983; Gartner and Stamatiadis 2002, 2009; Grafton and Newell 1967; Hunt et al. 1981; Little et al. 1982; Morgan and Little 1964; Newell 1982; Webster 1958) that there is a particular value of the signal capacity given by formula.

$$C_{cl} = q_{sat} T_G^{(eff)} / \vartheta, \quad (26)$$

where q_{sat} is the saturation flow rate; $T_G^{(eff)}$ is the effective duration of green signal phase; $\vartheta = T_G + T_Y + T_R$ is the duration of the cycle time of traffic signal; and T_G , T_Y , and T_R are durations of the green, yellow, and red phases of traffic signal, respectively, that are assumed to be time independent (the physical sense of q_{sat} and $T_G^{(eff)}$ has been explained in the review (Gartner and Stamatiadis 2009) as well as in Chap. 9 of the

book Kerner 2017b). We call C_{cl} (26) as classical signal capacity.

In the classical theory of traffic at the signal, it is assumed that traffic breakdown occurs at the signal when the average arrival flow rate at the signal $\bar{q}_{in}$ (called as the average arrival traffic rate on the approach (Gartner and Stamatiadis 2009)).

$$\bar{q}_{in} = \vartheta^{-1} \int_0^{\vartheta} q_{in}(t) dt \quad (27)$$

is larger than the classical signal capacity:

$$\bar{q}_{in} > C_{cl}. \quad (28)$$

In (27), $q_{in}(t)$ is the arrival flow rate at the signal that can be a complex function of time.

In the classical theory (see, e.g. Dion et al. 2004; Gartner 1983; Gartner and Stamatiadis 2002, 2009; Grafton and Newell 1967; Hunt et al. 1981; Little et al. 1982; Morgan and Little 1964; Newell 1982; Webster 1958), undersaturated traffic exists at the signal only then, when the average arrival flow rate at the signal $\bar{q}_{in}$ is smaller than the classical capacity, i.e., when.

$$\bar{q}_{in} < C_{cl}. \quad (29)$$

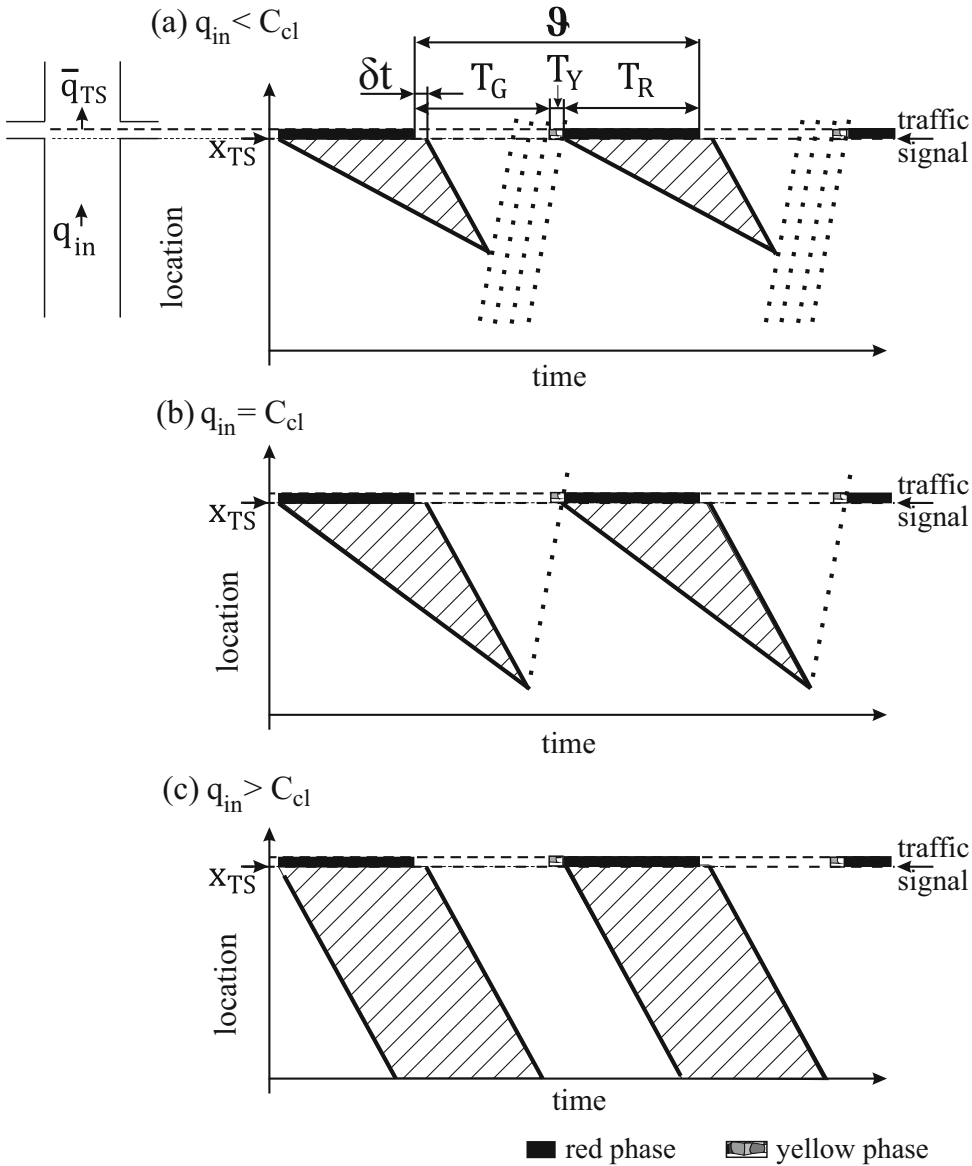
In this case, on average the queue has fully dissolved, and the last vehicle of this queue passes the signal location before the next red phase begins (Fig. 19a).

When fluctuations in the arrival flow rate and in the saturated flow rate are neglected, in the classical theory, it is assumed that in undersaturated traffic under condition.

$$\bar{q}_{in} = C_{cl}, \quad (30)$$

the rate of traffic passing the signal q_{TS} is fully determined by the queue discharge (Fig. 19b). Queue patterns shown by solid lines in Fig. 19 are related to a time-independent arrival flow rate q_{in} .

Under condition (28) there are on average vehicles at the queue end that must stop at the signal during the next red phase. This means that under condition (28), the queue grows, i.e., oversaturated traffic must occur at the signal (Dion



Breakdown in Traffic Networks, Fig. 19 Qualitative explanation of traffic breakdown at signal in classical theory of city traffic at a time-independent arrival flow rate q_{in} : (a–c) Queue patterns (solid lines) in undersaturated traffic at $q_{in} < C_{cl}$ (a), $q_{in} = C_{cl}$ (b) and in well-developed oversaturated traffic at $q_{in} > C_{cl}$ (c); dotted lines – vehicle trajectories. In (a), δt is the lost time (the physical sense of δt has been explained in the review (Gartner and Stamatiadis 2009) as well as in Chap. 9 of the book (Kerner 2017b)), x_{TS} is the signal location, and $\bar{q}_{TS}$ is the

average flow rate downstream of traffic signal. Results are taken from Dion et al. (2004), Gartner (1983), Gartner and Stamatiadis (2002, 2009), Grafton and Newell (1967), Hunt et al. (1981), Little et al. (1982), Morgan and Little (1964), Newell (1982), Webster (1958) (Figures have been modified from the originals presented in Dion et al. 2004; Gartner 1983; Gartner and Stamatiadis 2002, 2009; Grafton and Newell 1967; Hunt et al. 1981; Little et al. 1982; Morgan and Little 1964; Newell 1982; Webster 1958)

et al. 2004; Gartner 1983; Gartner and Stamatiadis 2002, 2009; Grafton and Newell 1967; Hunt et al. 1981; Little et al. 1982; Morgan and Little 1964;

Newell 1982; Webster 1958). Associated queue patterns at the signal location for well-developed oversaturated traffic are shown by solid lines in

Fig. 19c. These results of the classical theory are associated with the assumption that the saturation flow rate q_{sat} is the maximum rate of flow on a street (Gartner and Stamatiadis 2009).

In the classical theory, it is assumed that the presented theory of traffic breakdown at the signal is valid independent of the time dependence of the arrival flow rate $q_{\text{in}}(t)$: When the average arrival flow rate $\bar{q}_{\text{in}}$ exceeds the classical signal capacity C_{cl} (28), then the transition from undersaturated traffic to oversaturated traffic, i.e., the breakdown at the signal, does occur.

Metastability of Undersaturated Traffic at Signal

Contrarily to the classical theory of traffic breakdown presented in section “Classical Capacity of Traffic Signal,” in the theory of traffic breakdown developed by the author recently (Kerner 2011b, 2013b, 2014), traffic breakdown at the signal is a time-delayed transition from undersaturated traffic to oversaturated traffic at the signal (time-delayed traffic breakdown at the signal). It has also been found in this theory of traffic breakdown at the signal (Kerner 2011b, 2013b, 2014) that features of the transition from undersaturated traffic to oversaturated traffic at the signal depend qualitatively on the time dependence of the arrival flow rate $q_{\text{in}}(t)$.

It turns out that there is a finite range of the average arrival flow rate that exceeds the classical signal capacity (28) within which there can be two states of traffic at the signal:

- A metastable state of undersaturated traffic.
- A state of oversaturated traffic.

Due to the existence of the metastable undersaturated traffic, the transition from under- to oversaturated traffic at the signal exhibits the nucleation nature. Due to the nucleation nature of traffic breakdown at the signal, under condition (28) within a finite range of the average arrival flow rate discussed below, there is a time delay to the transition from under- to oversaturated traffic at the signal. In this section, we make an overview of general features of the time-delayed traffic breakdown at the signal (Kerner 2011b, 2013b,

2014) (a detailed consideration of the time-delayed traffic breakdown at the signal has been done in Chap. 9 of the book (Kerner 2017b)).

The general features of the time-delayed traffic breakdown at the signal are as follows Kerner (2011b, 2013b, 2014):

- (i) There should not be necessarily traffic breakdown at the signal in metastable undersaturated traffic at the signal: Under condition (28), within a finite range of the flow rate $\bar{q}_{\text{in}}$, a metastable state of undersaturated traffic can persist at the signal over a long time interval that can be considerably longer than the signal cycle ϑ .
- (ii) In the metastable state of the undersaturated traffic, traffic breakdown (transition from undersaturated traffic to oversaturated traffic) at the signal occurs after a time delay denoted by $T^{(B)}$.
- (iii) This time delay of traffic breakdown at the signal $T^{(B)}$ is a random value.
- (iv) The time delay to the breakdown $T^{(B)}$ can take many signal cycles.
- (v) The above-listed features of the time-delayed traffic breakdown at the signal mean that at the same given average arrival flow rate as well as given signal parameters under condition (28) within a finite range of the flow rate $\bar{q}_{\text{in}}$, metastable undersaturated traffic can exist during several signal cycles.

These features of traffic breakdown at the signal explain why we call the model of traffic breakdown at the signal as the *model of random time-delayed traffic breakdown at the signal*. The model of random time-delayed traffic breakdown at the signal introduced firstly in Kerner (2011b) has been further developed for different parameters of the arrival flow rate and traffic signal in Kerner (2013b, 2014), Kerner et al. (2014).

In Kerner (2013b, 2014), it has been shown that there are some features of random time-delayed traffic breakdown at the signal, which are general ones for many traffic flow models. These general features of random time-delayed transition from undersaturated to oversaturated traffic at the signal are as follows:

1. At any time instant, there is an infinite number of signal capacities C that are between a minimum signal capacity $C_{\min}$ and a maximum signal capacity $C_{\max}$:

$$C_{\min} \leq C \leq C_{\max}, \quad (31)$$

where $C_{\min} < C_{\max}$. The existence of an infinite number of signal capacities at any time instant means that signal capacity is stochastic.

2. The minimum signal capacity $C_{\min}$ is equal to the classical signal capacity C_{cl} (26):

$$C_{\min} = C_{cl}. \quad (32)$$

3. Within a range of the average arrival flow rate $\bar{q}_{in}$

$$C_{\min} \leq \bar{q}_{in} < C_{\max}, \quad (33)$$

undersaturated traffic is in a metastable state with respect to the transition from metastable undersaturated traffic to oversaturated traffic (traffic breakdown) at the signal: Any average arrival flow rate $\bar{q}_{in}$ that satisfies (33) is a signal capacity.

4. In metastable undersaturated traffic, the probability $P^{(B)}(\bar{q}_{in})$ of spontaneous traffic breakdown at the signal during the observation time T_{ob} is an increasing function of the average arrival flow rate $\bar{q}_{in}$ (Fig. 20).

5. There can be two different regions of the metastable undersaturated traffic within which traffic breakdown exhibits different probabilistic features:

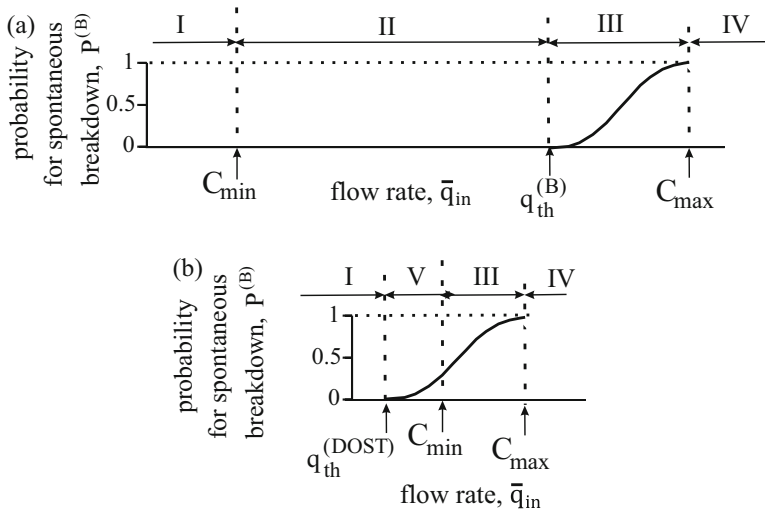
- Under condition

$$q_{th}^{(B)} \leq \bar{q}_{in} < C_{\max}, \quad (34)$$

during a chosen observation time T_{ob} , a time-delayed spontaneous traffic breakdown occurs with some probability (region III in Fig. 20a). In (34), $q_{th}^{(B)}$ is a threshold average arrival flow rate. Within this region III of the average arrival flow rate (Fig. 20a), the probability of spontaneous traffic breakdown at the signal during the observation time T_{ob} satisfies condition.

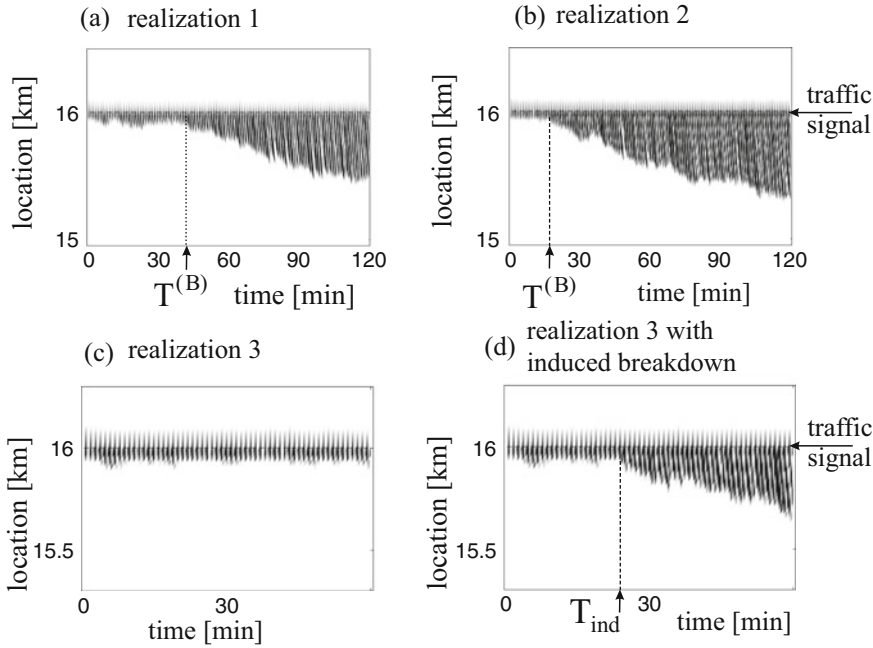
$$0 < P^{(B)}(\bar{q}_{in}) < 1. \quad (35)$$

This means that in region III of metastable undersaturated traffic (Fig. 20a, b) either



Breakdown in Traffic Networks, Fig. 20 Probabilistic characteristics of random time-delayed traffic breakdown at the signal: Two types (a) and (b) (see explanations in text) of qualitative dependencies of probability $P^{(B)}(\bar{q}_{in})$ of spontaneous traffic breakdown at the signal during the observation time T_{ob} as function of the average arrival

flow rate $\bar{q}_{in}$. The flow rate range I is related to stable undersaturated traffic, II – metastable undersaturated traffic, III – metastable undersaturated traffic in which spontaneous breakdown can occur, IV – unstable undersaturated traffic, and region V (in (b)) – dissolving oversaturated traffic (DOST) (Adapted from Kerner 2014)



Breakdown in Traffic Networks, Fig. 21 Simulations of different realizations of metastable undersaturated traffic under condition (28). Speed data in space–time plane are shown presented by regions with variable shades of gray

(shades of gray vary from white to black when the speed decreases from 50 km/h (white) to zero (black)). $\vartheta = 1$ min, $T^{(B)} = 40$ min (a) and 20 min (b) (Adapted from Kerner 2011b)

spontaneous traffic breakdown or induced traffic breakdown can occur (Fig. 21).

- Under condition

$$C_{\min} \leq \bar{q}_{\text{in}} < q_{\text{th}}^{(B)} \quad (36)$$

(region II in Fig. 20a), probability for spontaneous traffic breakdown at the signal during the observation time T_{ob} satisfies condition.

$$P^{(B)} = 0. \quad (37)$$

This means that in region II of metastable undersaturated traffic (Fig. 20a), only induced traffic breakdown is possible.

6. At $\bar{q}_{\text{in}} = C_{\max}$, traffic breakdown occurs spontaneously with probability $P^{(B)} = 1$ during the observation time T_{ob} . This means that under condition

$$\bar{q}_{\text{in}} \geq C_{\max}, \quad (38)$$

undersaturated traffic can be considered as unstable undersaturated traffic (region IV in Fig. 20).

7. Under condition

$$\bar{q}_{\text{in}} < C_{\min}, \quad (39)$$

features of a random time-delayed transition from undersaturated traffic to oversaturated traffic depend qualitatively on the time dependence of the arrival flow rate $q_{\text{in}}(t)$. The latter result explains two different dependencies of the breakdown probability $P^{(B)}$ on the average arrival flow rate $\bar{q}_{\text{in}}$ presented in Fig. 20a, b.

8. When the arrival flow rate $q_{\text{in}}(t)$ during the green signal phase is considerably larger than it is during the red signal phase, as this occurs often in a green wave (GW) in a city, then the function $P^{(B)}(\bar{q}_{\text{in}})$ is given by that shown in Fig. 20a. In this case, undersaturated traffic is stable under condition (39) (region I in Fig. 20a).

9. Contrarily to item 8, if the arrival flow rate $q_{in}(t)$ during the green and red signal phases is almost the same, then the function $P^{(B)}(\bar{q}_{in})$ is given by that shown in Fig. 20b. In the case shown in Fig. 20b, under condition (39) dissolving oversaturated traffic (DOST) can occur (region V in Fig. 20b). In dissolving oversaturated traffic, random emergence and subsequent random dissolution of a growing queue at traffic signal follow each other randomly.

Theoretical Fundamental of Transportation Science: Breakdown Minimization (BM) Principle

Real traffic and transportation networks consist of many road links. Each of the network links can have a bottleneck. Therefore, there can be a lot of network bottlenecks in a traffic network. At each of the network bottlenecks, traffic breakdown can occur, if the flow rate at the bottleneck is large enough. Obviously, each of the traffic control or optimization methods should try to prevent traffic breakdown in a traffic network.

In the standard dynamic traffic assignment and network optimization methodology, travel times or/and other travel costs on network links are considered to be self-evident traffic characteristics for objective functions used for optimization transportation problems. The main aim of associated generally accepted classical approaches is to minimize travel times or/and other travel costs in a traffic or transportation network. This standard methodology is related to the state of the art in traffic and transportation research (see, e.g., reviews and books (Bell and Iida 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Hoogendoorn et al. 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Taylor 2002)).

However, when the total network inflow rate increases, the minimization of travel times in the network based on the standard dynamic traffic

assignment methodology leads to considerably larger increases in the flow rates on short network routes (paths) in comparison with increases in the flow rates on long network routes. Due to the increase in the flow rates on short network routes, the probability of traffic breakdown can increase considerably on these short network routes. Therefore, traffic breakdown can randomly occur in the network. In other words, the minimization of travel times and/or other travel costs in a traffic network based on the state of the art in traffic and transportation research can deteriorate the traffic system considerably while provoking heavy traffic congestion in urban networks (see a brief discussion of this critical statement in section “Deterioration of Traffic in Networks Through Standard Dynamic Traffic Assignment” below and a detailed discussion in Chap. 13 of the book (Kerner 2017b)).

To prevent traffic breakdown in a traffic or transportation network, in 2011 the author introduced a breakdown minimization principle (BM principle) (Kerner 2011a). The application of the BM principle should lead to the minimization of the probability of traffic breakdown in a traffic or transportation network.

- The motivation for the BM principle is as follows. Methods for traffic control, dynamic traffic assignment, or other types of network optimization based on the BM principle should ensure the *minimum probability of traffic congestion* in the network.

Definition of the BM Principle

We consider a traffic or transportation network with N network bottlenecks, where $N > 1$. We assume that there are M network links (where $M > 1$) for which the inflow rates q_m , $m = 1, 2, \dots, M$ can be adjusted; q_m is the link inflow rate for a link with index m .

The breakdown minimization (BM) principle for dynamic traffic assignment, the optimization, and control in traffic and transportation networks is defined as follows Kerner (2011a):

- The BM principle states that the optimum of a traffic or transportation network with

N bottlenecks (where $N > 1$) is reached, when dynamic traffic assignment, optimization, and control are performed in the network in such a way that the probability P_{net} for the occurrence of traffic breakdown in at least one of the network bottlenecks during a given time interval T_{ob} for observing traffic flow reaches the minimum possible value.

- The BM principle is equivalent to the maximization of the probability that during the time interval T_{ob} traffic breakdown occurs at *none* of the network bottlenecks.

The objective of any application of the BM principle is to find the distinct assignment of network link inflow rates in a network and the set of control parameters of network bottlenecks at which the probability P_{net} that traffic breakdown occurs during the time interval T_{ob} in at least one of the network bottlenecks exhibits a minimum. For short, we call the probability P_{net} as the “probability of traffic breakdown in the network.”

In contrast with the standard dynamic traffic assignment and network optimization methodology that is the state of the art in traffic and transportation research (see section “[Deterioration of Traffic in Networks Through Standard Dynamic Traffic Assignment](#)”), the BM principle for dynamic traffic assignment, control, and optimization of traffic and transportation networks does *not* include travel times (or other “travel costs”) on different network routes (paths).

Constrain for Alternative Network Routes (Paths)

To avoid the use of network routes with considerably longer travel times in comparison with the shortest routes, the BM principle is applied for some “alternative network routes (paths)” (for short, alternative routes) *only*. This means that there is a constrain for alternative network routes (paths) by any application of the BM principle (constrain “alternative routes” for short). The constrain for the alternative routes prevents the use of too long routes in dynamic traffic assignment with the use of the BM principle.

Basic Applications of BM Principle

The objective of any application of the BM principle is to find the distinct assignment of network link inflow rates q_m in a network and the set of control parameters of network bottlenecks at which the probability of traffic breakdown in the network P_{net} exhibits the minimum possible value denoted by

$$P_{\text{net}} = P_{\text{net}}^{(\min)}. \quad (40)$$

At a small enough total network inflow rate Q , we can expect that no traffic breakdown can occur at the network bottlenecks. In this case, the minimum probability of traffic breakdown in the network $P_{\text{net}} = P_{\text{net}}^{(\min)} = 0$, i.e.,

$$P_{\text{net}} = 0. \quad (41)$$

A different situation can be expected, when the total network inflow rate Q is large enough. Then although the application of the BM principle reduces the probability of traffic breakdown in the network to the minimum possible value $P_{\text{net}} = P_{\text{net}}^{(\min)}$ (40), this minimum probability of traffic breakdown in the network $P_{\text{net}} = P_{\text{net}}^{(\min)} > 0$. This means that the probability of traffic breakdown in the network resulting from the application of the BM principle remains larger than zero:

$$P_{\text{net}} > 0. \quad (42)$$

In other words, there can be two different applications of the BM principle:

1. **The network throughput maximization approach related to case (41).** When the total network inflow rate Q is small enough, no application of the BM principle is required. However, when the total network inflow rate Q increases, there can be an application of the BM principle for dynamic traffic assignment in the network that keeps condition (41) while maximizing the network throughput. Such an application of the BM principle for “zero breakdown probability” revealed in Kerner (2016b) permits to find the distinct assignment of network link inflow rates q_m for the case

$P_{\text{net}} = 0$ (41). The application of the BM principle for “zero breakdown probability” (41) that maximizes the network throughput has been called *network throughput maximization approach* (Kerner 2016b, 2017a).

With the use of the network throughput maximization approach, general physical conditions for the maximization of the network throughput at which free flow conditions are ensured in the whole traffic or transportation network can be found. It turns out that there is a physical measure of the network called a *network capacity* that characterizes general features of the network with respect to the maximization of the network throughput (Kerner 2016b, 2017a).

2. **The application of the BM principle related to case (42).** This application of the BM principle is related to a large enough value of the total network inflow rate Q , when the probability of traffic breakdown in the network cannot be decreased to zero. The application of the BM principle permits the distinct assignment of network link inflow rates q_m in the network while minimizing the probability of traffic breakdown in the network (Kerner 2011a, b, 2016b, 2017a).

The BM principle can be applied only when there is free flow at bottlenecks of a network. Therefore, a question arises:

- Can the BM principle be applied when due to traffic breakdown at some of the bottlenecks in a network, there is traffic congestion in a part(s) of the network?

As introduced in Kerner (2011a, b), if traffic breakdown has already occurred at a network bottleneck, then network optimization with the BM principle can consist of the stages:

- (i) The dissolution of congestion through the use of congested pattern control approach at network bottlenecks (section “[Congested Pattern Control Approach Based on Three-Phase Theory](#)”).

- (ii) When the total network inflow rate Q is large enough, this congestion dissolution is not always possible to ensure. In this case, spatial limitation of congestion growth at network bottlenecks should be a target of congested pattern control approach.
- (iii) The minimization of traffic breakdown probability with the BM principle in the remaining network, i.e., the network part that is not influenced by congestion.

Network Throughput Maximization Approach

For each of the network bottlenecks $k = 1, 2, \dots, N$, the following probabilistic characteristics of traffic breakdown used below can be distinguished:

- A minimum capacity $C_{\min}^{(k)}$.
- The probability $P^{(B, k)}$ that during the time interval T_{ob} spontaneous traffic breakdown occurs at network bottleneck k , where $q_{\text{sum}}^{(k)}$ is the flow rate in free flow at bottleneck k .

At small enough values Q and $q_{\text{sum}}^{(k)}$, the following conditions are valid for each of the network bottlenecks:

$$q_{\text{sum}}^{(k)} < C_{\min}^{(k)}, \quad k = 1, 2, \dots, N. \quad (43)$$

Conditions (43) mean that free flow is stable with respect to traffic breakdown at each of the network bottlenecks. Therefore, no traffic breakdown can occur in the network.

When the total network inflow rate Q is small enough, no dynamic traffic assignment is needed. Due to the increase in Q over time, the flow rate $q_{\text{sum}}^{(k)}$ at least for one of the network bottlenecks $k = k_1^{(1)}$ becomes very close to $C_{\min}^{(k)}$. Therefore, in formula.

$$q_{\text{sum}}^{(k)} + \varepsilon^{(k)} = C_{\min}^{(k)} \quad \text{for } k = k_1^{(1)}, \quad (44)$$

a positive value $\varepsilon^{(k)} = C_{\min}^{(k)} - q_{\text{sum}}^{(k)}$ becomes very small:

$$\varepsilon^{(k)}/C_{\min}^{(k)} \ll 1 \text{ for } k = k_1^{(1)}. \quad (45)$$

When the total network inflow rate Q increases subsequently, we begin to use the network throughput maximization approach for dynamic traffic assignment in the network that is as follows Kerner (2016b). We maintain condition (44) at the expense of the increase in q_m on other alternative routes. As a result, condition $q_{\text{sum}}^{(k)} + \varepsilon^{(k)} = C_{\min}^{(k)}$ at $\varepsilon^{(k)}/C_{\min}^{(k)} \ll 1$ is satisfied for another bottleneck $k = k_2^{(1)}$. When Q increases further, we repeat the above procedure for other network bottlenecks. Consequently, we get

$$\begin{aligned} q_{\text{sum}}^{(k)} + \varepsilon^{(k)} &= C_{\min}^{(k)} \text{ for } k = k_z^{(1)}, \\ q_{\text{sum}}^{(k)} &< C_{\min}^{(k)} \text{ for } k \neq k_z^{(1)}, \\ z &= 1, 2, \dots, Z_1, \text{ where } Z_1 \geq 1, Z_1 \leq N; \\ k &= 1, 2, \dots, N; \end{aligned} \quad (46)$$

values $\varepsilon^{(k)}$ satisfy conditions.

$$0 < \varepsilon^{(k)}/C_{\min}^{(k)} \ll 1 \quad (47)$$

for bottlenecks $k = k_z^{(1)}$, $z = 1, 2, \dots, Z_1$. Thus, when the network inflow rate Q increases, we maintain conditions (46) at the expense of the increase in q_m on other alternative routes.

However, due to the constrain “alternative routes,” the number Z_1 of bottlenecks satisfying (46) is limited by some value Z (where $Z < N$). All values $\varepsilon^{(k)}$ in (46) accordingly to (47) are positive ones. Therefore, under conditions (46), conditions (43) are satisfied. Thus, due to the application of the network throughput maximization approach, free flow remains to be stable with respect to traffic breakdown at each of the network bottlenecks. For this reason, no traffic breakdown can occur in the network.

This means that conditions (46) in the limit case $Z_1 = Z$ and $\varepsilon^{(k)} \rightarrow 0$ are related to the *maximization of network throughput* at which free flow conditions are still ensured in the whole network. This is because at the subsequent increase in the total network inflow rate Q , the constrain “alternative routes” (section “[Constrain for Alternative](#)

[Network Routes \(Paths\)](#)”) does not allow us to maintain conditions (46) at the expense of the increase in q_m on other alternative routes: At least for one of the network bottlenecks, condition.

$$q_{\text{sum}} \geq C_{\min} \quad (48)$$

is satisfied. Therefore, traffic breakdown can occur in the network.

- The network throughput maximization approach is an approach to dynamic traffic assignment and control in a traffic or transportation network. The network throughput maximization approach maximizes the network throughput while keeping free flow conditions in the whole network. The network throughput maximization approach maintains condition $P_{\text{net}} = 0$, where P_{net} is the probability of traffic breakdown in the network. Therefore, the network throughput maximization approach is also called the BM principle for “zero breakdown probability.” Through the maximization of the network throughput with the network throughput maximization approach, no traffic breakdown can occur in the network.

A detailed explanation of the necessity of the application of the network throughput maximization approach for future dynamic traffic assignment and control in traffic and transportation networks can be found in Sect. 14.2 of the book (Kerner 2017b).

A Physical Measure of Traffic and Transportation Networks: Network Capacity

In many cases, the total network inflow rate $Q(t)$ changes in a traffic or transportation network over time very slowly in comparison with any characteristic times of dynamic traffic effects at network bottlenecks under free flow conditions in the network. For this reason, as it is usually assumed in classical theories of traffic and transportation networks, at any given Q free flow distribution in the

network can be considered as a steady state (steady state analysis of traffic and transportation networks) (Sheffi 1984). This means that it is assumed that $Q = Q_{\text{out}}$, where $Q_{\text{out}}(t)$ is the total network outflow rate.

Within a steady-state analysis of traffic and transportation networks, the limit case $Z_1 = Z$ in conditions (46) allows us to define a network measure (or “metric”) – *network capacity* denoted by C_{net} as follows Kerner (2016b). The network capacity C_{net} is the maximum total network inflow rate Q at which conditions

$$\begin{aligned} q_{\text{sum}}^{(k)} &= C_{\min}^{(k)} \text{ for } k = k_z^{(1)} \\ q_{\text{sum}}^{(k)} &< C_{\min}^{(k)} \text{ for } k \neq k_z^{(1)}, \\ z &= 1, 2, \dots, Z; \quad Z \geq 1, Z \leq N; \quad k = 1, 2, \dots, N \end{aligned} \quad (49)$$

are satisfied. From a comparison of conditions (46) and (49), we can see that the total network inflow rate reaches the network capacity

$$Q = C_{\text{net}} \quad (50)$$

when the limit case $Z_1 = Z$ in (46) is realized and all values $\varepsilon^{(k)}$ in (46) are set to zero: $\varepsilon^{(k)} = 0$. In accordance with the definition of the network capacity (49) and condition (48), at $Q = C_{\text{net}}$ (50) free flow becomes the metastable one with respect to traffic breakdown at network bottlenecks $k = k_z^{(1)}, z = 1, 2, \dots, Z$ (49). This means that under conditions (49), traffic breakdown can occur in the network.

When

$$Q < C_{\text{net}}, \quad (51)$$

as explained above, conditions (46) are satisfied. Under conditions (46), free flow is stable with respect to traffic breakdown in the whole network. Therefore, no traffic breakdown can occur in the network. This explains the physical sense of the network capacity C_{net} : Dynamic traffic assignment in the network in accordance with the network throughput maximization approach maximizes the network throughput at which traffic breakdown cannot occur in the whole network.

Respectively, when

$$Q > C_{\text{net}}, \quad (52)$$

due to the constrain “alternative routes” of the network throughput maximization approach at least for one of the network bottlenecks, the flow rate in free flow at the bottleneck $q_{\text{sum}}^{(k)}$ becomes larger than the minimum bottleneck capacity $C_{\min}^{(k)}$. Thus, rather than conditions (49) under condition (52), free flow is the metastable one with respect to traffic breakdown at the related network bottleneck(s). This means that under condition (52), traffic breakdown can occur in the network.

- The network capacity C_{net} is a measure (or “metric”) of a traffic or transportation network. The network capacity allows us to formulate a general physical condition for the maximization of the network throughput at which free flow does persist in the whole network: Under application of the network throughput maximization approach, as long as the total network inflow rate is smaller than the network capacity, traffic breakdown cannot occur in the network. The network capacity is a stochastic value.

Deterioration of Traffic in Networks Through Standard Dynamic Traffic Assignment

To find traffic optima and make an effective traffic control in traffic and transportation networks, a huge number of models for dynamic traffic assignment as well as other advanced traffic models have been introduced (see, e.g., reviews and books (Bell and Iida 1997; Daganzo 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Hoogendoorn et al. 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Taylor 2002)).

Most of the traffic researchers and practitioners strongly believe that the minimization of travel times and/or other travel costs in the network

should be the explicit objective of any reasonable principle for dynamic control, optimization, and assignment in traffic and transportation networks. In other words, travel times and other travel costs in a traffic or transportation network have been generally accepted in classical traffic and transportation theories to be self-evident traffic characteristics for objective functions used for the optimization of transportation networks, like dynamic traffic assignment (see, e.g., reviews and books (Bell and Iida 1997; Daganzo 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Hoogendoorn et al. 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Taylor 2002)).

The main aim of the classical approaches is to minimize travel times or/and other travel costs in the network. The minimization of network travel times (or/and other travel costs) is the state of the art in dynamic traffic assignment and control in traffic and transportation networks (see, e.g., reviews and books (Bell and Iida 1997; Daganzo 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Hoogendoorn et al. 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Taylor 2002)).

- The main objective of this section “Deterioration of Traffic in Networks Through Standard Dynamic Traffic Assignment” is to show that applications of this standard methodology of the minimization of network travel times deteriorate traffic system basically while provoking heavy traffic congestion in the network.

Classical Wardrop’s Equilibria

To show that applications of the standard methodology of the minimization of network travel times provoke heavy traffic congestion in a traffic or transportation network, we consider below the most famous approach to an analysis of traffic and transportation networks that is based on the

Wardrop’s user equilibrium (UE) and system optimum (SO) equilibrium introduced by Wardrop in 1952 (Wardrop 1952). The Wardrop’s UE and SO are also called the Wardrop’s principles or the Wardrop’s equilibria.

Wardrop’s UE: Traffic on a network distributes itself in such a way that the travel times on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel times.

Wardrop’s SO: The network-wide travel time should be a minimum.

The Wardrop’s principles reflect either the wish of drivers to reach their destinations as soon as possible (UE) or the wish of network operators to reach the minimum network-wide travel time (SO).

BM Principle Versus Wardrop’s Equilibria: General Results

We consider the network model of section “[Theoretical Fundamental of Transportation Science: Breakdown Minimization \(BM\) Principle](#).” We assume that a steady state analysis of a traffic or transportation network can be applied (see section “[Theoretical Fundamental of Transportation Science: Breakdown Minimization \(BM\) Principle](#)”). The steady state analysis of the networks is a standard analysis of most application of the Wardrop’s equilibria (see, e.g., reviews and books (Bell and Iida 1997; Daganzo 1997; Friesz and Bernstein 2016; Mahmassani 2001; Peeta and Ziliaskopoulos 2001; Rakha and Tawfik 2009; Ran and Boyce 1996; Sheffi 1984) as well as conference proceedings (Allsop et al. 2007; Cassidy and Skabardonis 2011; Ceder 1999; Hoogendoorn et al. 2013; Kuwahara et al. 2015; Lam et al. 2009; Lesort 1996; Mahmassani 2005; Taylor 2002)). Under a steady state of a network, we can apply the measure of the network called the network capacity C_{net} (section “[A Physical Measure of Traffic and Transportation Networks: Network Capacity](#)”). In this section “[BM Principle Versus Wardrop’s Equilibria: General Results](#),” we consider only the case when the total network inflow rate Q is smaller than the network capacity C_{net} , i.e., condition (51) is satisfied.

In other words, a general analysis of the BM principle versus the Wardrop's equilibria is performed below *only* for the case $Q < C_{\text{net}}$ (51). As we have shown, in this case the application of the BM principle for “zero breakdown probability, i.e., the network throughput maximization approach (46), ensures that traffic breakdown cannot occur in the network. One of the benefits of this analysis is that we can find a general explanation of the deterioration of traffic system through application of the Wardrop's equilibria that is independent of network characteristics.

In section “[Network Throughput Maximization Approach](#),” it has been shown that conditions (46) in the limit case $Z_1 = Z$ and $\varepsilon^{(k)} \rightarrow 0$ are related to the maximization of the network throughput at which free flow conditions are still ensured in the whole network: Free flow is stable with respect to traffic breakdown at each of the network bottlenecks, i.e., no traffic breakdown can occur in the network. In its turn, in the limit case $Z_1 = Z$ and $\varepsilon^{(k)} \rightarrow 0$ within a steady state analysis of the network, condition (51) is still valid that is used below.

A real traffic or transportation network consists of alternative routes with very different lengths. Therefore, at small enough link inflow rates q_m , there are routes with short travel times (“short routes”) and routes with longer travel times (“long routes”). When the total network inflow rate Q and, consequently, values q_m increase, the minimization of travel times in the network with the use of dynamic traffic assignment based on the Wardrop's UE or SO leads to considerably larger increases in the flow rates on short routes in comparison with increases in the flow rates on long routes (see, e.g., Rakha and Tawfik 2009; Sheffi 1984).

Thus, through the application of the Wardrop's UE or SO, the flow rate on a short route should be larger than the flow rate on a long route. This is independent of values $C_{\min}^{(k)}$ for bottlenecks on the route. Contrarily, in accordance with conditions (46) resulting from the network throughput maximization approach, the flow rate at any network bottleneck $q_{\min}^{(k)}$ should be smaller than the minimum capacity $C_{\min}^{(k)}$ of the bottleneck. This does not depend on whether a bottleneck is on a short route or the bottleneck is on a long route.

Therefore, at $Q \rightarrow C_{\text{net}}$ (however, the total network inflow rate is still smaller than the network capacity: $Q < C_{\text{net}}$ (51)), rather than conditions (46), any application of the Wardrop's UE or SO leads to conditions

$$q_{\min}^{(k)} < C_{\min}^{(k)} \quad (53)$$

for some of the bottlenecks on long routes, whereas for some of the bottlenecks on short routes, we get

$$q_{\min}^{(k)} > C_{\min}^{(k)}. \quad (54)$$

Condition (54) means that free flow becomes metastable one with respect to the breakdown at these bottlenecks. Therefore, in accordance with condition (54), traffic breakdown can occur at network bottlenecks on the short routes.

The above conclusion that at $Q \rightarrow C_{\text{net}}$ any application of the Wardrop's equilibria results in the occurrence of the metastable free flow at some of the network bottlenecks can additionally be explained through a consideration of the following hypothetical network: In the network with different lengths of alternative routes, values of the minimum capacity $C_{\min}$ are assumed the same for any network bottleneck. We assume that our statement about the metastability of free flow at some of the network bottlenecks might be incorrect: An application of the Wardrop's UE or SO might result in conditions (46) at which free flow is stable with respect to traffic breakdown at each of the network bottlenecks. However, at $C_{\min}^{(k)} = C_{\min}$ and $\varepsilon^{(k)} \rightarrow 0$ in (46), on network routes with different lengths, the flow rates $q_{\text{sum}}^{(k)}$ must be the same for all bottlenecks for which conditions (46) are satisfied. This contradicts the sense of any application of the Wardrop's UE or SO: On average, the flow rates should be larger on short routes than those on long routes.

Thus, due to the application of the Wardrop's UE or SO, already at $Q < C_{\text{net}}$ (51), it turns out that for some of the network bottlenecks $q_{\text{sum}}^{(k)} > C_{\min}^{(k)}$ (54). Therefore, traffic breakdown can occur at these bottlenecks. In this case, it is not possible to predict the time instant at which traffic breakdown

occurs at these bottlenecks. This is because the time delay to the breakdown $T^{(B)}$ is a random value (see Sect. 5.4.4 of the book Kerner 2017b).

After traffic breakdown has occurred at a network bottleneck, we can apply the Wardrop's UE or SO for dynamic traffic assignment without any delay. This is possible because the speed decrease due to the breakdown at the network bottleneck can be measured. However, after the assignment has been made, there is always a time delay in the change of traffic variables at the bottleneck location. This time delay is caused by travel time from the beginning of a link to the bottleneck location on the link. Therefore, it is not possible to avoid congested traffic occurring due to traffic breakdown at the network bottleneck. A control method can only effect on a spatiotemporal distribution of congestion in the network. From this general analysis, we can see that the main benefit of the network throughput maximization approach in comparison with the Wardrop's equilibria is as follows:

- The wish of humans to use shortest routes of a network contradicts fundamentally another wish of humans to drive under free flow conditions in the network. Therefore, the use of the classical Wardrop's equilibria results basically in the occurrence of congestion in urban networks. This can explain why approaches to dynamic traffic assignment based on the classical Wardrop's equilibria have failed by their applications in the real world.

Random Sequence of Phase Transitions in Network

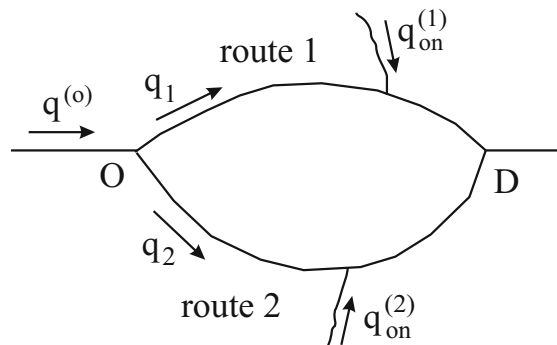
In a general consideration made in section “[BM Principle Versus Wardrop's Equilibria: General Results](#),” we have shown that the Wardrop's equilibria should deteriorate the traffic system while leading to traffic congestion. Here, for the case $Q < C_{\text{net}}$ (51), we illustrate this general conclusion for a simple model of two-route network shown in Fig. 22.

The Wardrop's UE for two-route network shown in Fig. 22 results in

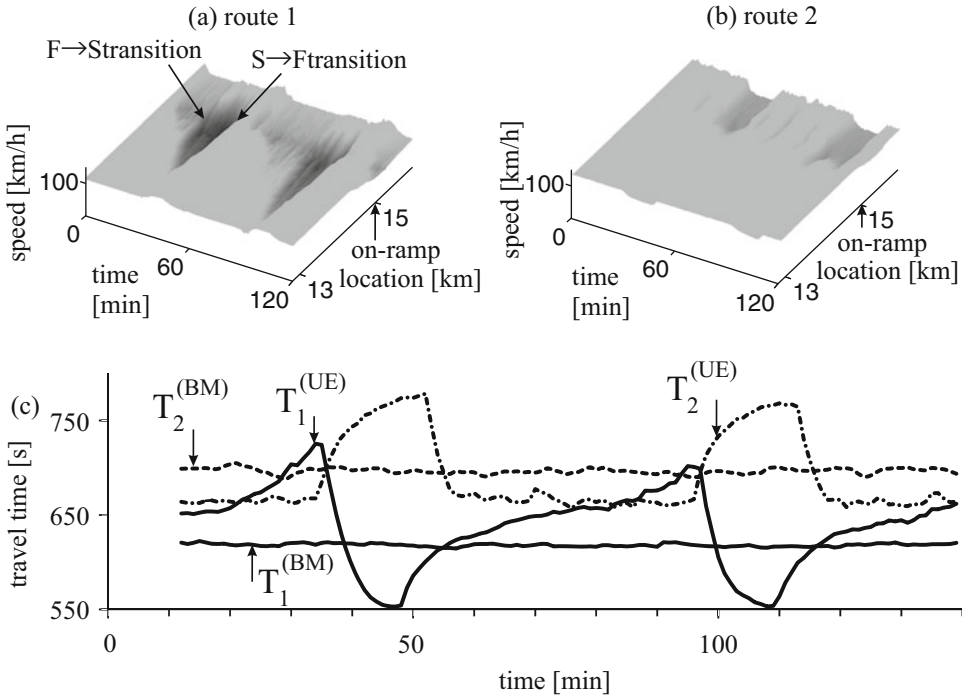
$$T_1(q_1, q_{\text{on}}^{(1)}) = T_2(q_2, q_{\text{on}}^{(2)}), \quad q^{(o)} = \sum_{m=1}^2 q_m, \quad (55)$$

where T_s , $s = 1, 2$ are travel times on routes $s = 1$ and $s = 2$, respectively.

In Fig. 23, we have applied dynamic traffic assignment of the flow rates q_1 and q_2 in (55) based on congested pattern control approach. In this dynamic feedback control method, bottleneck control starts *only* after a random traffic breakdown is registered by the road detector. In our case, after delay time $T_1^{(B)} = 20$ min, traffic breakdown has occurred at the on-ramp bottleneck on route 1 (labeled by “F $\rightarrow$ S transition” in Fig. 23a). A congested pattern resulting from the breakdown has been registered by the detector. Without any delay, the flow rate q_1 decreases to a value $q_1 - \Delta q$, and the flow rate q_2 increases to the value $q_2 + \Delta q$.



Breakdown in Traffic Networks, Fig. 22 Model of two-route network: Sketch of network with two routes and two on-ramp bottlenecks; $L_1 = 20$ km, $L_2 = 25$ km



Breakdown in Traffic Networks, Fig. 23 Simulations of dynamic traffic assignment for two-route network (Fig. 22) at time-independent inflow rate $q^{(o)}$. (a, b) Speed in space and time on routes 1 (a) and 2 (b) under application of the Wardrop's UE (55) and congested pattern control approach (see explanations in the text) at $\Delta q = 1600$ vehicles/h. (c) Travel times $T_1 = T_1^{(UE)}(t)$, $T_2 = T_2^{(UE)}(t)$ for the Wardrop's UE. Travel times $T_1 = T_1^{(BM)}$,

$T_2 = T_2^{(BM)}$ are related to application of the network throughput maximization approach (BM principle for "zero breakdown probability" $P_{\text{net}} = 0$ (41)) for which $q_1 = 3560$ vehicles/h, $q_2 = 2340$ vehicles/h. $Q = 7000$ vehicles/h. Arrows $F \rightarrow S$ and $S \rightarrow F$ show time instants of $F \rightarrow S$ and $S \rightarrow F$ transitions, respectively

In Fig. 23, we have used a value Δq at which congested patterns have been dissolved at the on-ramp bottleneck on route 1 and no traffic breakdown is realized at the on-ramp bottleneck on route 2. Thus, after traffic breakdown has occurred at the on-ramp bottleneck on route 1, due to the decrease in the flow rate on route 1 to $q_1 - \Delta q$, the congested pattern at the on-ramp bottleneck on route 1 has dissolved over time (labeled by "S $\rightarrow$ F transition" in Fig. 23a). As a result, free flow has returned at the on-ramp bottleneck on route 1. Therefore, in accordance with the control method, the flow rates q_1 and q_2 return to their values found from Eq. (55). Then, after another random time delay, traffic breakdown has occurred at the on-ramp bottleneck on route 1 once more (Fig. 23a). A congested pattern resulting from this second breakdown has been

registered by the detector. Without any delay, the flow rate q_1 decreases to a value $q_1 - \Delta q$, and the flow rate q_2 increases to the value $q_2 + \Delta q$. Over time the second congested pattern has dissolved (Fig. 23a), and free flow has returned at the on-ramp bottleneck on route 1 and so on.

- Because traffic breakdown occurs at the bottleneck after a random value of delay time $T^{(B)}$, the common result of the application of the Wardrop's UE principle for dynamic traffic assignment is a *random process* of the congested pattern emergence with the subsequent dissolution of the congested pattern (Fig. 23a). Respectively, we observe complex oscillations of route travel times $T_1^{(UE)}(t)$ and $T_2^{(UE)}(t)$ (Fig. 23c).

Conclusions: Future Dynamic Traffic Assignment and Control in Networks

Main conclusions of this entry are as follows:

1. The empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck can be considered an empirical fundamental of transportation science.
2. The three-phase theory is the theoretical fundamental of transportation science that explains the empirical fundamental of transportation science.
3. The BM principle is the theoretical fundamental of transportation science that minimizes traffic congestion in a traffic or transportation network.
4. There is a general approach for the maximization of the network throughput. This network maximization throughput approach allows us to find general conditions for the *maximization* of the network throughput at which traffic breakdown cannot occur in the whole network.
5. Although there are great achievements of classical traffic flow theories and models in the understanding of many important empirical traffic phenomena, the classical theories and traffic flow models cannot explain and show the empirical fundamental of transportation science – the nucleation nature of traffic breakdown at a highway bottleneck. The metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck has no sense for the classical traffic theories. This explains the failure of applications of the classical traffic and transportation theories and models in the real world.
6. Applications of the classical traffic flow models for an analysis of the effect of ITS (intelligent transportation systems) on traffic flow lead to invalid conclusions about the ITS performance in real traffic.
7. The three-phase theory is incommensurable with the classical traffic flow theories.

As emphasized above, the three-phase theory and the BM principle can be considered the theoretical fundamentals of transportation science.

Therefore, there are at least the following challenging areas for transportation science in which the three-phase theory and the BM principle should be used in the future:

1. The development of new traffic flow models in the framework of the three-phase theory should be intensified. This is because only three-phase traffic flow models that incorporate the metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck can be used for reliable analysis of the performance of ITS in traffic flow.
2. Congested pattern control approach based on the three-phase theory should be further developed and implemented to the market.
3. The three-phase theory should be also a theoretical basis for the future development of effective methods of traffic control in networks in the case of an extremely large total network inflow rate at which congestion propagates through the whole network. Indeed, in many real traffic networks, there are not enough alternative routes to avoid traffic congestion at large enough traffic demand. Thus, at large enough traffic demand, even “perfect” methods of traffic control and dynamic network optimization could not prevent traffic breakdown. Nevertheless, the application of ITS can change characteristics of traffic congestion with the objective to increase traffic safety and comfort while moving in congested traffic.
4. A further theoretical development of the network throughput maximization approach (the BM principle for the case “zero breakdown probability”) as well as other applications of the BM principle together with congested pattern control approach should be made.
5. Traffic control centers as well as traffic guidance systems in vehicles should be developed that permit dynamic traffic assignment in traffic networks based on the applications of the BM principle.
6. Studies of microscopic (single vehicle) empirical features of traffic breakdown at network bottlenecks and resulting congested patterns should be intensified. In particular, the

empirical studies should give the answer to questions about features of congested traffic that can effectively be influenced through the use of ITS-applications for the spatial limitation of congestion growth and/or for the dissolution of traffic congestion.

7. The development of assistant systems in vehicles as well as automated driving (self-driven) vehicles that can influence on spatial limitation of congestion growth and/or on congestion dissolution in traffic and transportation networks should be made.
8. The development of automated driving vehicles whose dynamic behavior is automatically adaptable to dynamic behavior of manual drivers should be performed.
9. The network throughput maximization approach should be applied for the design of future urban areas.
10. Features of the nucleation nature of traffic breakdown and applications of the BM principle for dynamic traffic assignment and control in traffic and transportation networks discussed in this entry together with spatio-temporal features of the propagation of traffic congestion should be used for development of realistic models of travel decision behavior and transportation demand.

Acknowledgments I would like to thank Sergey Klenov and Hubert Rehborn for the help and useful suggestions. I thank our partners for their support in the projects “UR: BAN – Urban Space: User oriented assistance systems and network management” and “MEC-View – Object detection for automated driving based on Mobile Edge Computing,” funded by the German Federal Ministry of Economic Affairs and Energy.

Bibliography

Primary Literature

- Allsop RE, Bell MGH, Heydecker BG (eds) (2007) *Transportation and traffic theory 2007*. Elsevier Science Ltd, Amsterdam
- Banks JH (1989) Freeway speed-flow-concentration relationships: more evidence and interpretations (with discussion and closure). *Transp Res Rec* 1225:53–60
- Banks JH (1990) Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering? *Transp Res Rec* 1287:20–28
- Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793–800
- Bell MGH, Iida Y (1997) *Transportation network analysis*. John Wiley & Sons, Incorporated, Hoboken
- Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. *Math Mod Meth App Sc* 12:1801–1843
- Bose A, Ioannou P (2003) Mixed manual/semi-automated traffic: a macroscopic analysis. *Transp Res C* 11:439–462
- Brilon W, Zurlinden H (2004) Kapazität von Straßen als Zufallsgröße. *Straßenverkehrstechnik* 4:164
- Brilon W, Geistefeld J, Regler M (2005a) Reliability of freeway traffic flow: a stochastic concept of capacity. In: Mahmassani HS (ed) *Transportation and traffic theory. Proceedings of the 16th international symposium on transportation and traffic theory*. Elsevier, Amsterdam, pp 125–144
- Brilon W, Regler M, Geistefeld J (2005b) Zufallscharakter der Kapazität von Autobahnen und praktische Konsequenzen – Teil 1. *Straßenverkehrstechnik* 3:136
- Brockfeld E, Kühne RD, Skabardonis A, Wagner P (2003) Toward benchmarking of microscopic traffic flow models. *Transp Res Rec* 1852:124–129
- Brockfeld E, Kühne RD, Wagner P (2005) Calibration and validation of simulation models. In: *Proceeding of the transportation research board 84th annual meeting*, TRB paper no. 05-2152. TRB, Washington, DC
- Cambridge Systematics, Inc (2001) *Twin Cities ramp meter evaluation – final report*. Cambridge Systematics, Inc., Oakland. Prepared for the Minnesota Department of Transportation
- Cassidy MJ, Skabardonis A (eds) (2011) *Papers selected for the 19th international symposium on transportation and traffic theory*. *Procedia Soc Behav Sci* 17:1–716
- Ceder A (ed) (1999) *Transportation and traffic theory*. In: *Proceedings of the 14th international symposium on transportation and traffic theory*. Elsevier Science Ltd, Oxford
- Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6:165–184
- Cheng H (2011) *Autonomous intelligent vehicles: theory, algorithms, and implementation*. Springer, Berlin
- Chien CC, Zhang Y, Ioannou PA (1997) Traffic density control for automated highway systems. *Automatica* 33:1273–1285
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
- Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
- Daganzo CF (1994) The cell-transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B* 28:269–287
- Daganzo CF (1995) The cell transmission model, part II: network traffic. *Transp Res B* 29:79–93
- Daganzo CF (1997) *Fundamentals of transportation and traffic operations*. Elsevier Science Inc., New York

- Davis LC (2004) Effect of adaptive cruise control systems on traffic flow. *Phys Rev E* 69:066110
- Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. *Physica A* 379:274–290
- Davis LC (2014) Nonlinear dynamics of autonomous vehicles with limits on acceleration. *Physica A* 405:128–139
- Dion F, Rakha H, Kang YS (2004) Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transp Res B* 38:99–122
- Elefteriadou L (2014) An introduction to traffic flow theory, Springer optimization and its applications, vol 84. Springer, Berlin
- Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Transp Res Rec* 1484:80–89
- Elefteriadou L, Kondyli A, Brilon W, Hall FL, Persaud B, Washburn S (2014) Enhancing ramp metering algorithms with the use of probability of breakdown models. *J Transp Eng* 140:04014003
- Friesz TL, Bernstein D (2016) Foundations of network optimization and games, Complex networks and dynamic systems, vol 3. Springer, New York/Berlin
- Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) *Traffic and granular flow* '01. Springer, Heidelberg
- Gartner NH (1983) OPAC: a demand-responsive strategy for traffic signal control. *Transp Res Rec* 906:75–81
- Gartner NH, Stamatiadis C (2002) Arterial-based control of traffic flow in urban grid networks. *Math Comput Model Pergamon* 35:657–671
- Gartner NH, Stamatiadis C (2009) Traffic networks, optimization and control of Urban. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9470–9500
- Gartner NH, Messer CJ, Rathi A (eds) (2001) *Traffic flow theory: a state-of-the-art report*. Transportation Research Board, Washington, DC
- Gazis DC (2002) *Traffic theory*. Springer, Berlin
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7:499–505
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
- Gipps PG (1981) Behavioral car-following model for computer simulation. *Transp Res B* 15:105–111
- Gipps PG (1986) A model for the structure of lane-changing decisions. *Transp Res B* 20:403–414
- Grafton RB, Newell GF (1967) Optimal policies for the control of an undersaturated intersection. In: Edie LC, Herman R, Rothery R (eds) *Vehicular traffic science*. American Elsevier Publishing Company, New York, pp 239–257
- Greenshields BD, Bibbins JR, Channing WS, Miller HH (1935) A study of traffic capacity. *Highw Res Board Proc* 14:448–477
- Haight FA (1963) *Mathematical theories of traffic flow*. Academic, New York
- Hall FL (1987) An interpretation of speed-flow-concentration relationships using catastrophe theory. *Trans Res A* 21:191–201
- Hall FL (2001) Traffic stream characteristics. In: Gartner NH, Messer CJ, Rathi A (eds) *Traffic flow theory: a state-of-the-art report*. Transportation Research Board, Washington DC, pp 2-1–2-36
- Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Transp Res Rec* 1320:91–98
- Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transp Res Rec* 1365:12–18
- Hegyi A, Bellemans T, De Schutter B (2017) Freeway traffic management and control. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) *Traffic and granular flow* '99. Springer, Heidelberg
- Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7:86–106
- Highway Capacity Manual (2000) National Research Council, Transportation Research Board, Washington, DC
- Highway Capacity Manual (2010) National Research Council, Transportation Research Board, Washington, DC
- Hoogendoorn SP, Luding S, Bovy PHL, Schreckenberg M, Wolf DE (eds) (2005) *Traffic and granular flow* '03. Springer, Heidelberg
- Hoogendoorn SP, Knoop VL, van Lint H (eds) (2013) 20th international symposium on transportation and traffic theory (ISTTT 2013). *Procedia Soc Behav Sci* 80:1–996
- Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT a traffic responsive method of coordinating signals. TRRL report no. LR 1014. Transportation and Road Research Laboratory, Crowthorne
- Ioannou PA, Chien CC (1993) Autonomous intelligent cruise control. *IEEE Trans Veh Technol* 42:657–672
- ISO 15622 (2010) *Intelligent transport systems – adaptive cruise control systems – performance requirements and test procedures*
- Kerner BS (1998a) Theory of congested traffic flow. In: Rysgaard R (ed) *Proceedings of the 3rd symposium on highway capacity and level of service*, vol 2. Road Directorate, Ministry of Transport, Denmark, pp 621–642
- Kerner BS (1998b) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
- Kerner BS (1999a) Congested traffic flow: observations and theory. *Transp Res Rec* 1678:160–167
- Kerner BS (1999b) Theory of congested traffic flow: self-organization without bottlenecks. In: Ceder A (ed) *Transportation and traffic theory*. Elsevier Science, Amsterdam, pp 147–171

- Kerner BS (1999c) The Physics of Traffic. Phys World 12:25–30
- Kerner BS (2004a) The physics of traffic. Springer, Berlin/New York
- Kerner BS (2004b) Verfahren zur Ansteuerung eines in einem Fahrzeug befindlichen verkehrsadaptiven Assistenzsystems, German patent publication DE 10308256A1. <https://google.com/patents/DE10308256A1>; Patent WO 2004076223A1 (2004) <https://google.com/patents/WO2004076223A1>; EU Patent EP 1597106B1 (2006); German patent DE 502004001669D1 (2006)
- Kerner BS (2005) Control of spatiotemporal congested traffic patterns at highway bottlenecks. Physica A 355:565–601
- Kerner BS (2007a) Control of spatiotemporal congested patterns at highway bottlenecks. IEEE Trans ITS 8:308–320
- Kerner BS (2007b) On-ramp metering based on three-phase theory – part III. Traffic Eng Control 48:68–75
- Kerner BS (2007c) Three-phase traffic theory and its applications for freeway traffic control. In: Inweldi PO (ed) Transportation research trends. Nova Science Publishers, Inc., New York, pp 1–97
- Kerner BS (2007d) Method for actuating a traffic-adaptive assistance system which is located in a vehicle, USA patent US 20070150167A1. <https://google.com/patents/US20070150167A1>; USA patent US 7451039B2 (2008)
- Kerner BS (2007e) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assistenzsystem, German patent publication DE 102007008253A1. <https://register.dpma.de/DPMAREGISTER/pat/PatSchrifteneinsicht?doId=DE102007008253A1>
- Kerner BS (2007f) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assistenzsystem, German patent publication DE 102007008257A1. <https://register.dpma.de/DPMAREGISTER/pat/PatSchrifteneinsicht?doId=DE102007008257A1>
- Kerner BS (2008) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assistenzsystem, German patent publication DE 102007008254A1
- Kerner BS (2009) Introduction to modern traffic flow theory and control. Springer, Berlin/New York
- Kerner BS (2011a) Optimum principle for a vehicular traffic network: minimum probability of congestion. J Phys A Math Theor 44:092001
- Kerner BS (2011b) Physics of traffic gridlock in a city. Phys Rev E 84:045102(R)
- Kerner BS (2013a) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. Physica A 392:5261–5282
- Kerner BS (2013b) The physics of green-wave breakdown in a city. Europhys Lett 102:28010
- Kerner BS (2014) Three-phase theory of city traffic: moving synchronized flow patterns in under-saturated city traffic at signals. Physica A 397:76–110
- Kerner BS (2015a) Microscopic theory of traffic-flow instability governing traffic breakdown at highway bottlenecks: growing wave of increase in speed in synchronized flow. Phys Rev E 92:062827
- Kerner BS (2015b) Failure of classical traffic flow theories: a critical review. Elektrotechnik und Informationstechnik 132:417–433
- Kerner BS (2016a) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. Physica A 450:700–747
- Kerner BS (2016b) The maximization of the network throughput ensuring free flow conditions in traffic and transportation networks: breakdown minimization (BM) principle versus Wardrops equilibria. Eur Phys J B 89:199
- Kerner BS (2017a) Breakdown minimization principle versus Wardrops equilibria for dynamic traffic assignment and control in traffic and transportation networks: a critical mini-review. Physica A 466:626–662
- Kerner BS (2017b) Breakdown in traffic networks: fundamentals of transportation science. Springer, Berlin/New York
- Kerner BS (2017c) Traffic breakdown, modeling approaches to. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer, Berlin
- Kerner BS (2017d) Physics of Autonomous Driving based on Three-Phase Traffic Theory. arXiv:1710.10852v3 (<http://arxiv.org/abs/1710.10852>)
- Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. Phys Rev E 68:036130
- Kerner BS, Klenov SL (2009) Traffic breakdown, probabilistic theory of. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer, Berlin, pp. 9282–9303
- Kerner BS, Konhäuser P (1994) Structure and parameters of clusters in traffic flow. Phys Rev E 50:54–83
- Kerner BS, Rehborn H (1996a) Experimental features and characteristics of traffic jams. Phys Rev E 53: R1297–R1300
- Kerner BS, Rehborn H (1996b) Experimental properties of complexity in traffic flow. Phys Rev E 53: R4275–R4278
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. Phys Rev Lett 79:4030–4033
- Kerner BS, Klenov SL, Schreckenberg M (2014) Traffic breakdown at a signal: classical theory versus the three-phase theory of city traffic. J Stat Mech 2014:P03001
- Kometani E, Sasaki T (1958) On the stability of traffic flow (report-I). J Oper Res Soc Jpn 2:11
- Kometani E, Sasaki T (1959) A safety index for traffic with linear spacing. Oper Res 7:704
- Kometani E, Sasaki T (1961) Dynamic behavior of traffic with a nonlinear spacing. In: Herman R (ed) Theory of traffic flow. Elsevier, Amsterdam, pp 105–119
- Krauß S, Wagner P, Gawron C (1997) Metastable states in a microscopic model of traffic flow. Phys Rev E 55:5597–5602
- Kuhn TS (2012) The structure of scientific revolutions, 4th edn. The University of Chicago Press, Chicago/London
- Kukuchi S, Uno N, Tanaka M (2003) Impacts of shorter perception-reaction time of adapted cruise controlled

- vehicles on traffic flow and safety. *J Transp Eng* 129:146–154
- Kuwahara M, Kita H, Asakura Y (eds) (2015) 21st international symposium on transportation and traffic theory. *Transp Res Procedia* 7:1–704
- Lam WHK, Wong SC, Lo HK (eds) (2009) *Transportation and traffic theory 2009*. Springer, Dordrecht/Heidelberg/London/New York
- Lesort J-B (ed) (1996) *Transportation and traffic theory*. In: *Proceedings of the 13th international symposium on transportation and traffic theory*. Elsevier Science Ltd, Oxford
- Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
- Levine W, Athans M (1966) On the optimal error regulation of a string of moving vehicles. *IEEE Trans Automat Contr* 11:355–361
- Levinson D, Zhang L (2006) Ramp meters on trial: evidence from the twin cities metering holiday. *Transp Res A* 40:810–828
- Liang C-Y, Peng H (1999) Optimal adaptive cruise control with guaranteed string stability. *Veh Syst Dyn* 32:313–330
- Liang C-Y, Peng H (2000) String stability analysis of adaptive cruise controlled vehicles. *JSME Int J Ser C* 43:671–677
- Lighthill MJ, Whitham GB (1995) On kinematic waves. I flow movement in long rivers. II a theory of traffic flow on long crowded roads. *Proc Roy Soc A* 229:281–345
- Little JD, Kelson MD, Gartner NH (1982) MAXBAND: a program for setting signals on arterials and triangular networks. *Transp Res Rec* 795:40–46
- Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. *Trans Res Cir E-C018*:84–95
- Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419:1–64
- Mahmassani HS (2001) Dynamic network traffic assignment and simulation methodology for advanced system management applications. *Netw Spat Econ* 1:267–292
- Mahmassani HS (ed) (2005) *Transportation and traffic theory*. In: *Proceedings of the 16th international symposium on transportation and traffic theory*. Elsevier, Amsterdam
- May AD (1990) *Traffic flow fundamentals*. Prentice-Hall, Inc., Englewood Cliffs
- McDonald M, Wu J (1997) The integrated impacts of autonomous intelligent cruise control on motorway traffic flow. In: *Proceedings of the ISC 97 conference*, Boston
- Morgan JT, Little JDC (1964) Synchronizing traffic signals for maximal bandwidth. *Oper Res* 12:896–912
- Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys (France)* I 2:2221–2229
- Nagel K, Wagner P, Woessler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–716
- Newell GF (1961) Nonlinear effects in the dynamics of car following. *Oper Res* 9:209–229
- Newell GF (1963) Instability in dense highway traffic, a review. In: *Proceedings of the second international symposium on traffic road traffic flow*. OECD, London, pp 73–83
- Newell GF (1982) *Applications of queuing theory*. Chapman Hall, London
- Newell GF (2002) A simplified car-following theory: a lower order model. *Transp Res B* 36:195–205
- Papageorgiou M (1983) *Application of automatic control concepts in traffic flow modeling and control*. Springer, Berlin/New York
- Papageorgiou M, Kotsialos A (2000) Freeway ramp metering: an overview. In: *Proceedings of the 3rd annual IEEE conference on intelligent transportation systems (ITSC 2000)*, Dearborn, Oct 2000, pp 228–239
- Papageorgiou M, Kotsialos A (2002) Freeway ramp metering: an overview. *IEEE Trans Intell Transp Syst* 3(4):271–280
- Papageorgiou M, Papamichail I (2008) Overview of traffic signal operation policies for ramp metering. *Transp Res Rec* 2047:28–36
- Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990a) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: part I: modelling. *Transp Res Part A* 24A(5):345–359
- Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990b) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: part II: coordinated on-ramp metering. *Transp Res Part A* 24A(5):361–370
- Papageorgiou M, Hadj-Salem H, Blosseville J-M (1991) ALINEA: a local feedback control law for on-ramp metering. *Transp Res Rec* 1320:58–64
- Papageorgiou M, Hadj-Salem H, Middelham F (1997) ALINEA local ramp metering – summary of field results. *Transp Res Rec* 1603:90–98
- Papageorgiou M, Diakaki C, Dinopoulou V, Kotsialos A, Wang Y (2003) Review of road traffic control strategies. In: *Proceedings of the IEEE*, vol 91, pp 2043–2067
- Papageorgiou M, Wang Y, Kosmatopoulos E, Papamichail I (2007) ALINEA maximizes motorway throughput – an answer to flawed criticism. *Traffic Eng Control* 48(6):271–276
- Payne HJ (1971) *Models of freeway traffic and control*. In: Bekey GA (ed) *Mathematical models of public systems*, vol 1. Simulation Council, La Jolla
- Payne HJ (1979) FREFLO: A macroscopic simulation model of freeway traffic. *Transp Res Rec* 772:68
- Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. *Netw Spat Econ* 1:233–265
- Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Transp Res Rec* 1634:64–69
- Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24:274–287
- Prigogine I (1961) A Boltzmann-like approach to the statistical theory of traffic flow. In: Herman R (ed) *Theory of traffic flow*. Elsevier, Amsterdam, pp 158–164
- Rajamani R (2012) *Vehicle dynamics AML control*, Mechanical engineering series. Springer, Boston

- Rakha H, Tawfik A (2009) Traffic networks: dynamic traffic routing, assignment, and assessment. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp. 9429–9470
- Ran B, Boyce D (1996) Modeling dynamic transportation networks. Springer, Berlin
- Rehborn H, Klenov SL (2009) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp. 9500–9536
- Rehborn H, Klenov SL, Koller M (2017) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin
- Reuschel A (1950) Fahrzeugbewegungen in der Kolonne. *Österreichisches Ingenieur-Archiv* 4:193–215
- Richards PI (1956) Shockwaves on the highway. *Oper Res* 4:42–51
- Saifuzzaman M, Zheng Z (2014) Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transp Res C* 48:379–403
- Sathiyar SP, Kumar SS, Selvakumar AI (2013) A comprehensive review on cruise control for intelligent vehicles. *Int J Innov Technol Explor Eng (IJITEE)* 2:89–96
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) Traffic and granular flow' 05. In: *Proceedings of the international workshop on traffic and granular flow*. Springer, Berlin
- Schadschneider A, Chowdhury D, Nishinari K (2011) *Stochastic transport in complex systems*. Elsevier Science Inc., New York
- Schreckenberg M, Wolf DE (eds) (1998) Traffic and granular flow' 97. In: *Proceedings of the international workshop on traffic and granular flow*. Springer, Singapore
- Sheffi Y (1984) *Urban transportation networks: equilibrium analysis with mathematical programming methods*. Prentice-Hall, Englewood Cliffs
- Shladover SE (1995) Review of the state of development of advanced vehicle control systems (AVCS). *Veh Syst Dyn* 24:551–595
- Shvetsov VI (2003) Mathematical modeling of traffic flows. *Autom Remote Control* 64:1651–1689
- Smaragdis E, Papageorgiou M (2003) Series of new local ramp metering strategies. *Transp Res Rec* 1856:74–86
- Swaroop D, Hedrick JK (1996) String stability for a class of nonlinear systems. *IEEE Trans Autom Control* 41:349–357
- Swaroop D, Hedrick JK, Choi SB (2001) Direct adaptive longitudinal control of vehicle platoons. *IEEE Trans Veh Technol* 50:150–161
- Taylor MAP (ed) (2002) *Transportation and traffic theory in the 21st century*. In: *Proceedings of the 15th international symposium on transportation and traffic theory*. Elsevier Science Ltd, Amsterdam
- Treiber M, Helbing D (2001) Microsimulations of freeway traffic including control measures. *Automatisierungstechnik* 49:478–484
- Treiber M, Kesting A (2013) *Traffic Flow dynamics*. Springer, Berlin
- van Arem B, van Driel C, Visser R (2006) The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Trans Intell Transp Syst* 7:429–436
- Wardrop JG (1952) Some theoretical aspects of road traffic research. In: *Proceedings of the institute of civil engineers part II*, vol 1, pp 325–378
- Webster FV (1958) *Traffic signal settings*. HMSO, London
- Whitham GB (1974) *Linear and nonlinear waves*. Wiley, New York
- Wiedemann R (1974) *Simulation des Verkehrsflusses*. University of Karlsruhe, Karlsruhe
- Wolf DE (1999) Cellular automata for traffic simulations. *Physica A* 263:438–451
- Wolf DE, Schreckenberg M, Bachem A (eds) (1995) Traffic and granular flow. In: *Proceedings of the international workshop on traffic and granular flow*. World Scientific, Singapore
- Xiao L, Gao F (2010) A comprehensive review of the development of adaptive cruise control systems. *Veh Syst Dyn* 48:1167–1192
- Zurlinden H (2003) *Ganzjahresanalyse des Verkehrsflusses auf Straßen*. Schriftenreihe des Lehrstuhls für Verkehrswesen der Ruhr-Universität Bochum, Heft 26, Bochum

Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks

Hesham Rakha and Aly Tawfik
Center for Sustainable Mobility, Virginia Tech,
Transportation Institute, Virginia Polytechnic
Institute and State University, Blacksburg, VA,
USA

Article Outline

Glossary
Definition of the Subject
Introduction
Driver Travel Decision Behavior Modeling
Static Traffic Routing and Assignment
Dynamic Traffic Routing
Traffic Modeling
Dynamic Travel Time Estimation
Dynamic or Time-Dependent Origin-Destination
Estimation
Dynamic Estimation of Measures of Effectiveness
Use of Technology to Enhance System
Performance
Related Transportation Areas
Future Directions
Appendix
Bibliography

Glossary

Car-following model A mathematical representation (traffic flow model) for driver longitudinal motion behavior.

Dynamic traffic assignment Traffic assignment considering the temporal dimension of the problem.

Link or arc A roadway segment with homogeneous traffic and roadway characteristics (e.g. same number of lanes, base lane capacity,

free-flow speed, speed-at-capacity, and jam density). Typically networks are divided into links for traffic modeling purposes.

Marginal link travel time The increase in a link's travel time resulting from an assignment of an additional vehicle to this link.

Road pricing Road pricing is an economic concept in which drivers are charged for the use of the road facility.

Route or path A sequence of roadway segments (links or arcs) used by a driver to travel from his/her point of origin to his/her destination.

Static traffic assignment Traffic assignment ignoring the temporal dimension of the problem.

Synthetic O-D estimation The procedure that estimates O-D demands from measured link flow counts, which includes static, time-dependent, and dynamic.

System optimum traffic assignment The assignment of traffic such that the average journey travel times of all motorists is a minimum, which implies that the aggregate vehicle-hours spent in travel is also minimum.

Time-dependent traffic assignment An approximate approach to modeling the dynamic traffic assignment problem by dividing the time horizon into steady-state time intervals and applying a static assignment to each time interval.

Traffic assignment The procedure used to find the link flows from the Origin-Destination (O-D) demand. Traffic assignment involves two steps: (1) traffic routing and (2) traffic demand loading. Traffic assignment can be divided into static, time-dependent, and dynamic.

Traffic loading The procedure of assigning O-D demands to routes.

Traffic routing The procedure that computes the sequence of roadways that minimize some utility objective function. This utility function could either be travel time or a generalized function that also includes road tolls.

Traffic stream motion model A mathematical representation (traffic flow model) for traffic stream motion behavior.

User equilibrium traffic assignment The assignment of traffic on a network such that it distributes itself in a way that the travel costs on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel costs.

Definition of the Subject

The dynamic nature of traffic networks is manifested in both temporal and spatial changes in traffic demand, roadway capacities, and traffic control settings. Typically, the underlying network traffic demand builds up over time at the onset of a peak period, varies stochastically during the peak period, and decays at the conclusion of the peak period. As traffic congestion builds up within a transportation network, drivers may elect to either cancel their trip altogether, alter their travel departure time, change their mode of travel, or change their route of travel. Dynamic traffic routing is defined as the process of dynamically selecting the sequence of roadway segments from a trip origin to a trip destination. Dynamic routing entails using time-dependent roadway travel times to compute this sequence of roadway segments. Consequently, the modeling of driver routing behavior requires the estimation of roadway travel times into the near future, which may entail some form of traffic modeling.

In addition to dynamic changes in traffic demand, roadway capacities are both stochastic and vary dynamically as vehicles interact with one another along roadway segments. For example, the roadway capacity at a merge section varies dynamically as the composition of on-ramp and freeway demands vary (Cassidy and Rudjanakanoknad 2005; Cassidy and Windover 1995; Elefteriadou et al. 2005; Evans et al. 2001; Kerner 2004a, b, 2005; Kerner and Klenov 2006; Kerner et al. 2004; Lertworawanich and Elefteriadou 2001, 2003; Lorenz and Elefteriadou 2001; Minderhoud and Elefteriadou 2003; Rakha et al. 2004a). To further complicate matters, traffic

control settings (e.g. traffic signal timings) also vary both temporally and spatially, thus introducing another level of dynamics within transportation networks. All these factors make the dynamic assessment of traffic networks extremely complex, as shall be demonstrated in this article. The article is by no means comprehensive but does provide some insight into the various challenges and complexities that are associated with the assessment of dynamic networks.

Introduction

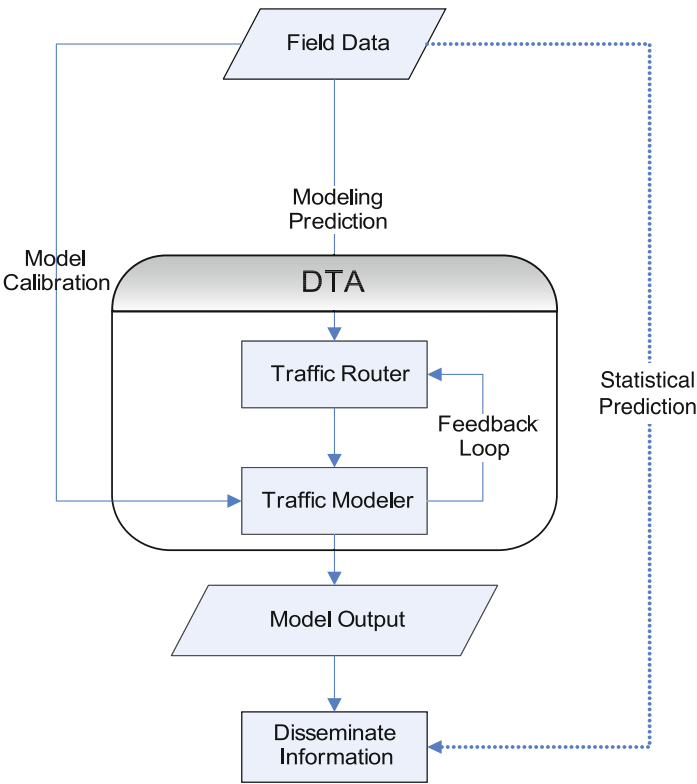
Studies have shown that even drivers familiar with a trip typically choose sub-optimal routes thus incurring extra travel time in the range of seven percent on average (Jeffery 1981). Furthermore, the occurrence of incidents and special events introduces other forms of variability that drivers are unable to anticipate and thus result in additional errors in a driver's route selection. Consequently, advanced traveler information systems (ATISs), which are an integral component of intelligent transportation systems (ITSs), can assist the public in their travel decisions by providing real-time travel information via route guidance systems; variable message signs (VMSs); the radio, or the web. It is envisioned that better travel information can enhance the efficiency of a transportation system by allowing travelers to make better decisions regarding their time of departure, mode of travel, and/or route of travel. An integral component of an ATIS is a dynamic traffic assignment (DTA) system. A DTA system predicts the transportation network state over a short time horizon (typically 15- to 60-min time horizon) by modeling complex demand and supply relationships through the use of sophisticated models and algorithms. The DTA requires two sets of input, namely demand and supply data. Demand represents the demand for travel and is typically in the form of mode-specific time-dependent origin-destination (O-D) matrices. Alternatively, the supply component models the movement of individual vehicles along a roadway typically using roadway specific speed-flow-density relationships together

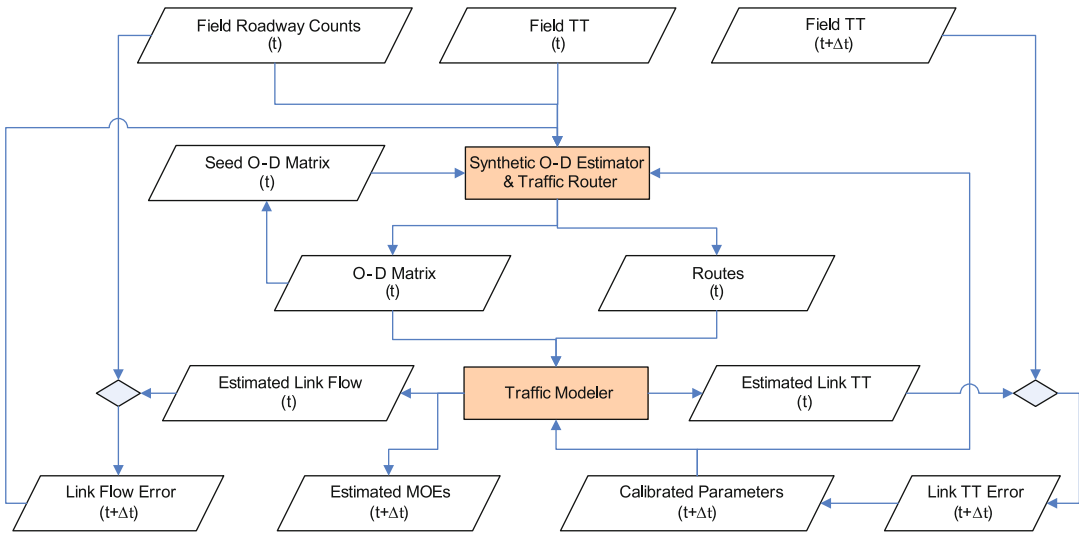
with the explicit modeling of queue buildup and decay. Figure 1 illustrates schematically that an ATIS can utilize two approaches for the estimation of future traffic conditions, namely: statistical models or a DTA framework. This article focuses on the DTA approach and thus will be described in more detail. The DTA combines a traffic router and modeler, as illustrated in the figure. The traffic router estimates the optimum travel routes while the traffic modeler models traffic to evaluate the performance of traffic after assigning motorists to their routes. A feedback loop allows for the feedback of either travel times or marginal travel times, which in turn, are used by the traffic router to compute the optimum routes. This feedback continues until the travel times are consistent with the travel routes and there is no incentive for drivers to alter their routes.

A DTA can be applied off-line (in a laboratory) or online (in the field). An on-line application of a DTA entails gathering traffic data in real-time at any instant t and feeding these data to the DTA to predict short-term traffic conditions Δt temporal

units into the future (i.e. at time $t + \Delta t$). As was mentioned earlier, the input to the DTA includes mode specific time dependent O-D matrices. Unfortunately, current surveillance equipment does not measure O-D matrices; instead they measure traffic volumes passing a specific point. Consequently, O-D estimation tools are required to estimate the O-D matrix from observed link counts, as illustrated in Fig. 2. However, the estimation of an O-D matrix requires identifying which O-D demands contribute to which roadway link counts. The assigning of O-D demands to link counts involves what is commonly known in the field of traffic engineering as the traffic assignment problem. Traffic assignment in turn requires real-time O-D matrices and roadway travel times as input. Consequently, some form of feedback is required to solve this problem. A more detailed description of traffic assignment formulations and techniques is provided in sections “Static Traffic Routing and Assignment” and “Dynamic Traffic Routing,” while the estimation of route travel times is described in section “Dynamic Travel

Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks,
Fig. 1 Schematic of an ATIS framework





Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks, Fig. 2 Dynamic traffic assessment and routing framework

[Time Estimation](#)” and the estimation of O-D matrices is described in section [“Dynamic or Time-Dependent Origin-Destination Estimation.”](#)

The dynamic assessment of traffic networks using a DTA is both data driven (trapezoidal boxes) and model based (colored rectangular boxes), as illustrated in Fig. 2. This procedure involves: measuring raw field data, constructing model input data, executing a traffic model to predict future conditions, and advising a traveler in the case of control systems. The framework starts by measuring traffic states at instant “ t ” (roadway travel times and link flows) and subsequently estimating these traffic states Δt in the future. Procedures for the estimation of dynamic roadway travel times are provided in section [“Dynamic Travel Time Estimation”](#) of this article. Using the measured link flows and travel times, an O-D matrix is constructed using a synthetic O-D estimator. Section [“Dynamic or Time-Dependent Origin-Destination Estimation”](#) describes the various formulations for estimating a dynamic O-D matrix together with some heuristic practical approaches to estimate this O-D matrix.

Once the O-D demands are estimated the future states are predicted using a traffic modeler. Section [“Traffic Modeling”](#) provides a brief overview of the various state-of-the-practice modeling

approaches. The model also computes various measures of effectiveness (MOEs) including delay, fuel consumption, and emissions, as will be described in section [“Dynamic Estimation of Measures of Effectiveness.”](#) The traffic modeler can either combine traffic modeling with traffic assignment or alternatively utilize the routes computed by the O-D estimator to route traffic. This closed loop optimal control framework can involve a single loop or in most cases may involve an iterative loop to attain equilibrium. The framework involves a feedback loop in which input model parameters are adjusted in real-time through the computation of an error between model predictions and actual measurements. This real-time calibration entails adjusting roadway parameters (e.g. capacity, free-flow speed, speed-at-capacity, and jam density) and traffic routes to reflect dynamic changes in traffic and network conditions. For example, the capacity of a roadway might vary because of changes in weather conditions and/or the occurrence of incidents. The system should be able to adapt itself dynamically without any user intervention.

This article attempts to synthesize the literature on the dynamic assessment and routing of traffic. The problem as will be demonstrated later in the paper is extremely complex because,

after all, it deals with the human psychic, which not only varies from one person to another, but may also vary depending on the purpose of a trip, the level of urgency the driver has, and the psychic of the driver at the time the trip is made. This article is by no means comprehensive, given the massive literature on the topic, but does highlight some of the key aspects of the problem, how researchers have attempted to address this problem, and future research needs and directions.

The article discusses the various issues associated with the dynamic assessment of transportation systems. Initially, driver travel decision behavior modeling is presented and discussed. Subsequently, various traffic assignment formulations are presented together with the implementation issues associated with these formulations. Next, the mathematical formulations of these assignment techniques are discussed together with mathematical and numerical approaches to modeling dynamic traffic routing. Subsequently, the issues associated with the modeling of traffic stream behavior, the estimation of dynamic roadway travel times, and the estimation of dynamic O-D demands are discussed. Subsequently, the procedures for computation of various assessment measures are presented. Next, the use of technology to alter driver behavior is presented. Finally, directions for further research are presented.

Driver Travel Decision Behavior Modeling

As with the general case of modeling human behavior, modeling driver travel behavior has always been complicated, never accurate enough, and in constant demand for further research. Among the early attempts to model human choice behavior is the economic theory of the “economic man”; who in the course of being economic is also “rational” (Simon 1955). According to Simon’s exact words, “*actual human rationality-striving can at best be an extremely crude and simplified approximation to the kind of global rationality that is implied, for example, by game-theoretical models.*”

In general, traffic assignment (static or dynamic assignment) has undoubtedly been among the most researched transportation problems, if not the most, for more than the past half of a century. However, DTA in particular has had the bigger share for almost one third of a century now. Since the early work of Merchant and Nemhauser (Merchant and Nemhauser 1978a, b), researchers have attempted to improve available DTA models, hence, providing a very rich and vastly wide literature.

As a result of the rapid technological evolution over the last decade of the previous century (the twentieth century); manifested in the communications, information and computational technological advances; a worldwide initiative to add information and communications technology to transport infrastructure and vehicles, termed as the intelligent transportation systems (ITS) program, was introduced to the transportation science. According to the Wikipedia Encyclopedia, among the main objectives of ITS is to “*manage factors that are typically at odds with each other such as vehicles, loads, and routes to improve safety and reduce vehicle wear, transportation times and fuel consumption.*” Needless to say, the ITS impact on route selection and roadway travel times has a direct effect on a DTA.

The main effect of ITS on DTA manifests itself within the area of advanced traveler information systems (ATIS). ATIS is primarily concerned with providing people, in general, and trip makers, in particular, with pre-trip and en-route trip-related information. According to the US Federal Highway Administration (FHWA), “*advanced traveler information includes static and real-time information on traffic conditions, and schedules, road and weather conditions, special events, and tourist information. ATIS is classified by how and when travelers receive their desired information (pre-trip or en-route) and is divided by user service categories. Operations essential to the success of these systems are the collection of traffic and traveler information, the processing and fusing of information often at a central point, and the distribution of information to travelers. Important components of these systems include new technologies applied to the use and presentation of*

information and the communications used to effectively disseminate this information” (Noonan and Shearer 1998).

As will be discussed later, a significant amount of DTA research is directed towards developing data dissemination standards. These standards attempt to achieve the maximum possible benefits while complying with the ITS objectives. Although the provision of pre-trip information may influence traveler departure time and route of travel (and in extreme cases, might result in a person canceling his/her trip all together), thus requiring further complicated DTA models that capture forgone and induced demand, as will be discussed later. Moreover, probably the greatest dimension for DTA model complexity was introduced to research when the disseminated ATIS information was to be designed as a control factor to change the manner by which trips are distributed over the network, for example from user equilibrium to system optimum.

Although ITS and ATIS were practically introduced a little more than a decade ago, and in spite of the significant research funds and efforts that have been devoted to the topic, current available DTA models are, at least, relatively undeveloped, which necessitates new approaches that can capture the challenges from the application domains as well as for the fundamental questions related to tractability and realism (Peeta and Ziliaskopoulos 2001). This will be discussed briefly in the following section.

Driver travel decision theory is a complicated research area. Research within this area encompasses a very wide range of research efforts. Before going over a brief list of these possible research areas, it should be noted that most of these research areas overlap with one another. Therefore, for a valid driver behavior model, all of the following aspects should be efficiently covered in a practical and realistic manner. This been said, the following is a brief list of some of the main research areas that are highly related to driver travel decision theory:

- Human decision theory, which can be reflected in the trip maker’s decision to make or cancel a scheduled trip, route and departure time

selection, compliance with the pre-trip or en-route disseminated information, en-route path diversion and/or return, mode choice based on disseminated information, etc. Literature concerning human decision theory extends back to more than half a century ago and continues to be researched up to this date. Examples of the literature concerning the human decision theory include: administrative behavior (Simon 1947, 1957), theory of choice (Arrow 1951), rational choice theory (Simon 1955), game theory, and decision field theory (Busemeyer and Townsend 1993). Examples of the literature concerning driver decision theory include: decision field theory (Talaat and Abdulhai 2006), approximate reasoning models (Koutsopoulos et al. 1995), route choice utility models (Hawas 2004), inductive learning (Nakayama and Kitamura 2000), effect of age on routing decisions (Walker et al. 1997), and rational learning (Nakayama et al. 2001).

- Design of disseminated information, which encompasses the criteria governing the dissemination of information, the structure and type of information to be disseminated, when data are disseminated, and identifying target drivers. This governs, to a large extent, the drivers’ compliance rates in response to disseminated information. Hence, affecting the routes chosen by drivers, the traffic volumes on these routes and alternative routes, and different travel times, among others. Literature concerning the effect of ATIS and ATIS content on drivers behavior include: the required information that would reduce traffic congestion (Arnott et al. 1991), the effect of ATIS on drivers route choice (Abdel-Aty et al. 1997), commuters diversion propensity (Schofer 1993), the effect of traffic information disseminated through variable message signs on driver choices (Peeta and Ramos 2006), drivers en-route routing decisions (Khattak et al. 1993).
- Human perception based on experience and information provision, which is reflected in day-to-day variations in driver decisions. For example, given identical conditions on two

separate days, the same person might select different routes and departure times; possibly due to different experiences on previous days. Examples of current literature include: models that include the incorporation of driver behavior dynamics under information provision (Peeta and Yu 2004), behavioral-based consistency seeking models (Peeta and Yu 2006), perception updating and day-to-day travel choice dynamics with information provision (Jha et al. 1998), the modeling of inertia and compliance mechanisms under real-time information (Srinivasan and Mahmassani 2000), drivers psychological deliberation while making dynamic route choices (Talaat and Abdulhai 2006), the effect of using in-vehicle navigational systems on driver behavior (Allen et al. 1991), the effect of network familiarity on routing decisions (Lotan 1997), and the effect of varying levels of cognitive loads on driver behavior (Katsikopoulos et al. 2000).

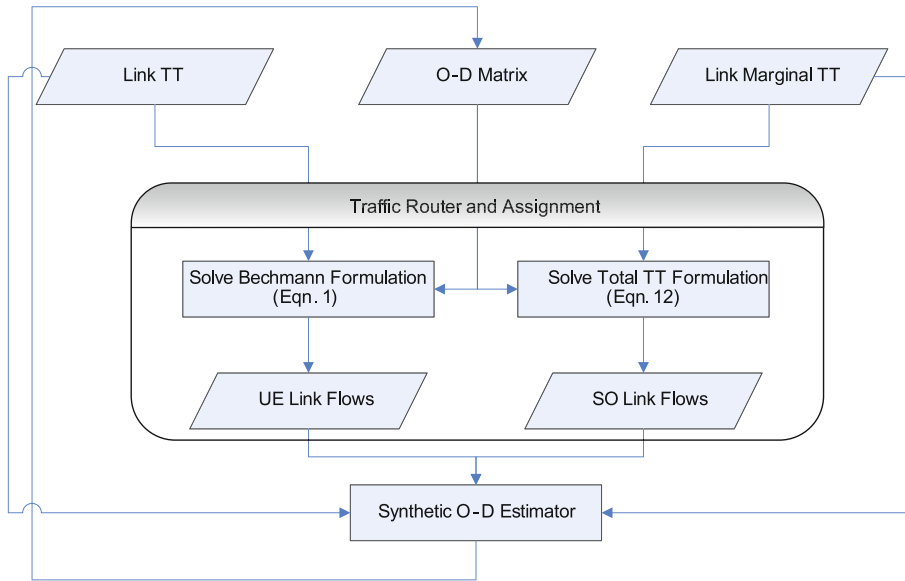
- Among the challenges in modeling human decision theory are the possible data collection techniques. The current practice for data collection includes revealed and stated preference surveys. Research has demonstrated that surveyed stated preference results have significant biases; in comparison to real behavior. In addition to the research being performed to analyze, capture, and improve the reasons for such biases; other research directions are being performed to solve other survey problems. For example, the problems of low and slow survey participation rates, as well as underrepresented groups in typical survey techniques. Examples of literature within this field include: stated preference for investigating commuters diversion propensity (Schofer 1993), using stated preference for studying the effect of advanced traffic information on drivers route choice (Abdel-Aty et al. 1997), driver response to variable message sign-based traffic information according to stated preference data collected through three different survey administration methods, namely, an on-site survey, a mail-back survey and an internet-based survey (Peeta and Ramos 2006),

transferring insights into commuter behavior dynamics from laboratory experiments to field surveys (Mahmassani and Jou 2000; Peeta et al. 2000), and the applicability of using driving simulators for data collection (Koutsopoulos et al. 1995).

- Issues of uncertainty, which is a fundamental feature in most transportation phenomena. Research dealing with uncertainty has a wide application in DTA. It can be represented in the trip maker route travel time estimates, in the compliance rates of drivers to information, in the driver's trust in the disseminated information and its reliability, among others. Uncertainty-related research issues have been addressed through several approaches, like stochastic modeling, fuzzy control, and reliability indices. Examples of current literature include the works of Birge and Ho (1993), Peeta and Zhou (1999a, b), Cantarell and Cascetta (1995), Ziliaskopoulos and Waller (2000), Waller and Ziliaskopoulos (2006), Waller (2000), Peeta and Yu (2006), Jha et al. (1998), Peeta and Paz (2006), Koutsopoulos et al. (1995), and Hawas (2004).

Static Traffic Routing and Assignment

Prior to describing the issues associated with dynamic routing, a description of static routing issues is first presented. This section describes two formulations for static traffic assignment, namely the User Equilibrium (UE) and System Optimum (SO) assignment. Traffic assignment is defined as the basic problem of finding the link flows given an origin-destination trip matrix and a set of link or marginal link travel times, as illustrated in Fig. 3. The solution of this problem can either be based on the assumption that each motorist travels on the path that minimizes his/her travel time – known as the UE assignment – or alternatively to minimize the system-wide travel time – known as the SO assignment. The traffic assignment initially computes the travel routes (paths) and then determines the unique link flows on the various network links. As will be discussed later, while the estimated link flows are unique the path



Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks, Fig. 3 Traffic assignment framework

flows that are derived from these link flows are not unique and thus require some computational tool to estimate the most-likely of these path flows (synthetic O-D estimator). If a time dimension is introduced to the assignment module the formulation is extended from a static to a dynamic context. However, as will be discussed later the addition of a time dimension deems the formulation non-convex and thus the mathematical program used to solve the problem becomes infeasible and thus comes the need for a simulation-based solution approach.

Wardrop (1952) was the first to explicitly differentiate between these two alternative traffic assignment methods or philosophies. Models based on Wardrop's first principle are referred to as UE, while those based on the second principle are deemed as SO. Wardrop's first principle states that "traffic on a network distributes itself in such a way that the travel costs on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel costs." Alternatively, Wardrop's second principle states that the average journey travel times of all motorists is a minimum, which implies that the aggregate vehicle-hours spent in travel is also minimum.

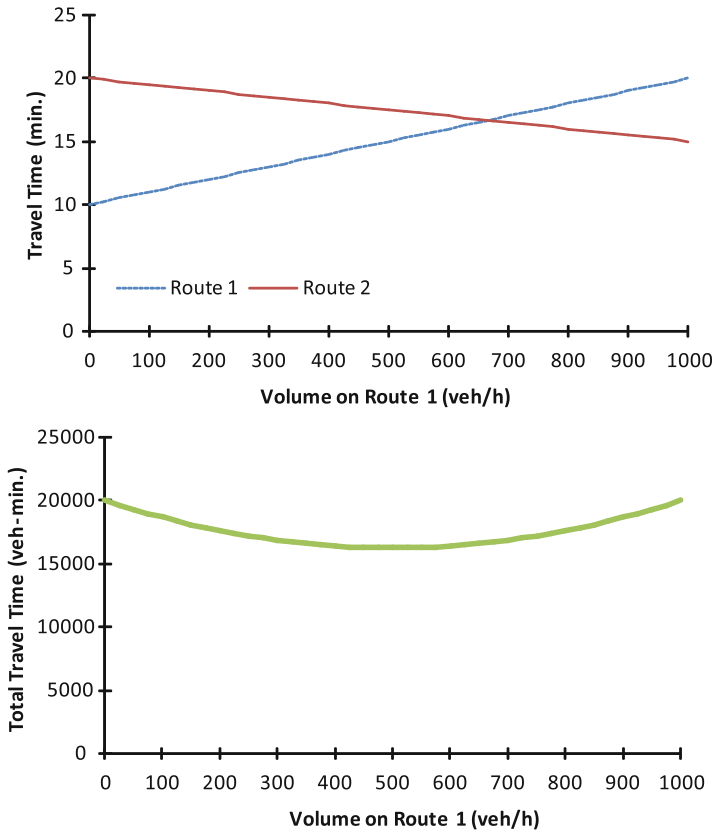
One of the most spectacular examples that illustrated that the UE flow in a network is in general different from the SO flow, is the Braess network (Braess 1968). In this network the system-optimal flow was obtained by completely suppressing the flow which would normally occur, on a certain link, at equilibrium. The Braess "paradox" was studied later in more detail (Fisk 1979; Frank 1981; LeBlanc 1975; LeBlanc and Abdulaal 1970; Murchland 1970; Rilett and Van Aerde 1991a; Steinberg and Zangwill 1983; Stewart 1980). For example, Stewart (Stewart 1980) illustrated three important facts using a very simple two-link network and the Braess paradox that included: (a) the equilibrium flow does not necessarily minimize the total cost; (b) adding a new link to a network may increase the total cost at equilibrium; (c) adding a new link to a network may increase the equilibrium travel cost for each individual motorist. Stewart also illustrated that a group of travelers having only one reasonable route may be seriously inconvenienced by another group of travelers who choose the same route in order to obtain a slight improvement in their personal cost of travel.

User Equilibrium Versus System Optimum
Traffic Assignment

The differences between user and system optimum traffic assignment are best illustrated using an example illustration. The sample test network for this study is derived from an earlier study by Rakha (1990). The network consists of two one-way routes, numbered 1 and 2, from origin A to destination B. The travel time relationship for route 1 is characterized by the relationship $10 + 0.010 v_1$ where v_1 is the traffic volume on route 1 (veh). Alternatively, the travel time along route 2 is characterized by the relationship $15 + 0.005 v_2$ where v_2 is the traffic volume traveling along route 2 (veh). Considering at total demand of 1000 veh traveling between zones A and B, the travel time along routes 1 and 2 vary as a function of the volume on each of the routes, as illustrated in Fig. 4. The figure demonstrates that the travel times along routes 1 and 2 are equal at 16.5 min when

667 veh travel along route 1 and 333 veh travel along route 2. Alternatively, the system-optimum traffic assignment is achieved at a volume distribution of 500 veh on routes 1 and 2, respectively. From a traffic engineering point of view, the difference in total travel time between the system and user-optimum traffic assignment (16,250 versus 16,667 veh-min) is of interest. This difference represents the extent of possible benefits for a system versus user optimum routing for this particular network and traffic pattern. Figure 4 also illustrates how the average link travel times on routes 1 and 2 vary for the same range of possible routings of traffic between route 1 and 2. In this figure the difference between the travel times on route 1 and 2 (15.0 versus 17.5 min) represents the incentive that exists for vehicles on route 2 to change to route 1. When compared to the user equilibrium routing, the total difference in travel time is composed of two components, which represent the respective

Dynamic Traffic Routing,
Assignment, and
Assessment of Traffic
Networks,
Fig. 4 Variation in route
and system travel time for
test network



increases (route 1) and decreases (route 2) in average travel time that result from a shift from the system to user-optimum routing.

Implementation Issues

While the simple example illustrated the potential benefits of system optimized routings and the incentive that exists for drivers to switch back to the original user equilibrium routings, it is clear that neither an exhaustive enumeration nor an analytical approach (solving the differential equations of the system travel time) are satisfactory for finding the system optimized routings when more than just a few possible routes are available.

Different static traffic assignment algorithms have been developed over the past half century. These methods are broadly divided into non-equilibrium and equilibrium methods. Non-equilibrium methods include all-or-nothing assignment, where all traffic is assigned to a single minimum path between two zones (path that incurs the minimum travel time). Example algorithms for computing minimum paths include models developed by Dantzig (1957) and Dijkstra (1959). Other non-equilibrium methods include incremental, iterative, diversion models, multipath assignment (Dial 1971), and combined models. According to Van Vliet (1976) the incremental assignment method (explained later) is capable of reaching an acceptable degree of convergence faster than an iterative method. With regards to diversion models, the most common diversion models include the California diversion curves (Moskowitz 1956) and the Detroit diversion curves (Smock 1962). Alternatively, multipath traffic assignment methods assign traffic stochastically. For example, the Dial method (Dial 1971) stochastically diverts trips to alternate paths, but trips are not explicitly assigned to routes. Other multipath methods (Burrell 1968, 1976) assume that users do not know the actual travel times on each link, but a driver's estimate of link travel time is drawn randomly from a distribution of possible times. Finally, combined non-equilibrium models include combining capacity restraint models with probabilistic assignment (Randle 1979), combining iterative with incremental assignment (Yagar 1971, 1974, 1975,

1976), or combining stochastic with equilibrium assignment (Sheffi and Powell 1981).

Equilibrium assignment techniques are based on Wardrop's first principle (Wardrop 1952). These were classified by Matsoukis and Michalopoulos (1986) into: assignments with fixed demand, assignments with elastic demands, and combined models. Only the first method will be discussed. The equilibrium assignment algorithm is a weighted combination of a sequence of all-or-nothing assignments. This produces a non-linear programming (NLP) problem which is subject to linear constraints. This NLP is very hard to solve and the approach seems to be of limited use for realistically sized equilibrium traffic assignment problems. The NLP problem can be replaced by a much simpler linear approximation and solved using the Frank–Wolfe algorithm (Frank and Wolfe 1956). This iterative linearization procedure still involves longer computational times than the iterative procedure. LeBlanc et al. (1974) developed an iterative procedure solving one-dimensional searches and LP problems that minimize successively better linear approximations to the non-linear objective function. Nguyen (1969) converted the convex optimization problem into a set of simpler sub-problems that could be solved with the convex-simplex method.

One of the most common approaches to implement a user equilibrium traffic assignment involves the use of an incremental traffic assignment technique (Leonard et al. 1978; Matsoukis 1986; Van Aerde 1985; Van Aerde and Yagar 1988a, b; Yagar 1971, 1975). Such a technique breaks down the total traffic demand that is to be loaded onto the network into a number of increments that are each loaded onto the network in turn. Each increment is loaded onto what appears to be the shortest route, after all the previous increments have been loaded. The link travel times are then recalculated, in order to re-compute the fastest route for the next increment to be loaded. When more than one route are to be used for travel between a given origin and destination, the increments are automatically assigned alternatively to each route, when each becomes faster again after previous increments head along the other route. In the end, the extent to which the overall assignment approaches an equilibrium

state depends upon the number of increments utilized, with the average final error being roughly proportional to the final increment size.

Van Aerde and Rakha (Rakha 1990; Rakha et al. 1989) demonstrated that the system-optimum traffic assignment can also be solved considering an incremental traffic assignment. Specifically, Van Aerde and Rakha (Rakha 1990; Rakha et al. 1989) recognized the fact that the increase in system travel time caused by the addition of one vehicle is composed of the additional travel time incurred by the subject vehicle and the increase in travel time that is imparted on all other vehicles which are already on the link. While the former quantity is usually already available as a direct or indirect measurement on the link, the derivation of the latter quantity is more subtle. It is a function of the rate of change of the average travel time, per additional vehicle, and the number of vehicles already on the link. In mathematical terms, this is simply the product of the derivative of the travel time versus volume relationship, with respect to volume, multiplied by the volume already present on the link. Consequently, the standard objective function that is utilized in any minimum path algorithm, which searches for the user equilibrium routes, can be replaced by a new objective function that minimizes the total travel time. This routing can be achieved using an incremental assignment of vehicles based on their marginal travel time as opposed to their actual travel time, which results in a system optimum as opposed to a user equilibrium routing, as was demonstrated earlier in Fig. 3. Stated differently at dynamic system optimum, the time-dependent marginal cost on all the paths actually used are equal and less than the marginal cost on any unused paths. In the static case, the path marginal cost (PMC) is the sum of the link marginal cost (LMC). However, in the dynamic case, the PMC evaluation is much more complicated since path flows are not assigned to links on the path simultaneously. However, within the dynamic context, most researchers (Ghali and Smith 1995; Peeta 1994) assume that the path flow perturbation travels along the path at the same speed as the additional flow unit. Shen et al. (2006) demonstrated that this assumption is not necessarily correct. Furthermore, they presented a solution algorithm for path-based system optimum models based on a new PMC

evaluation method. The approach was then tested and validated on a simple network.

Dynamic Traffic Routing

This section describes the mathematical formulations for the static routing problem together with some solution approaches to the problem. Subsequently, the extension of the problem for the dynamic context is presented together with state-of-the-art solution approaches.

The dynamic traffic assignment approach is summarized in Fig. 5 and involves three input variables, namely: dynamic link travel times (in the case of the UE assignment), dynamic marginal travel times (in the case of the SO assignment), and dynamic O-D matrices. In the case of the UE assignment the Bechmann formulation is solved Eq. (1) if we use a time-dependent static (or quasi static) assignment as will be discussed in detail in the following sections, while in the case of the SO assignment Eq. (12) is solved. Within the static context these formulations are solved analytically using a mathematical program given that the objective function and feasible region are convex. Alternatively, in the dynamic context the objective function is non-convex and thus is more difficult to solve necessitating the use of a modeling approach to solve the problem.

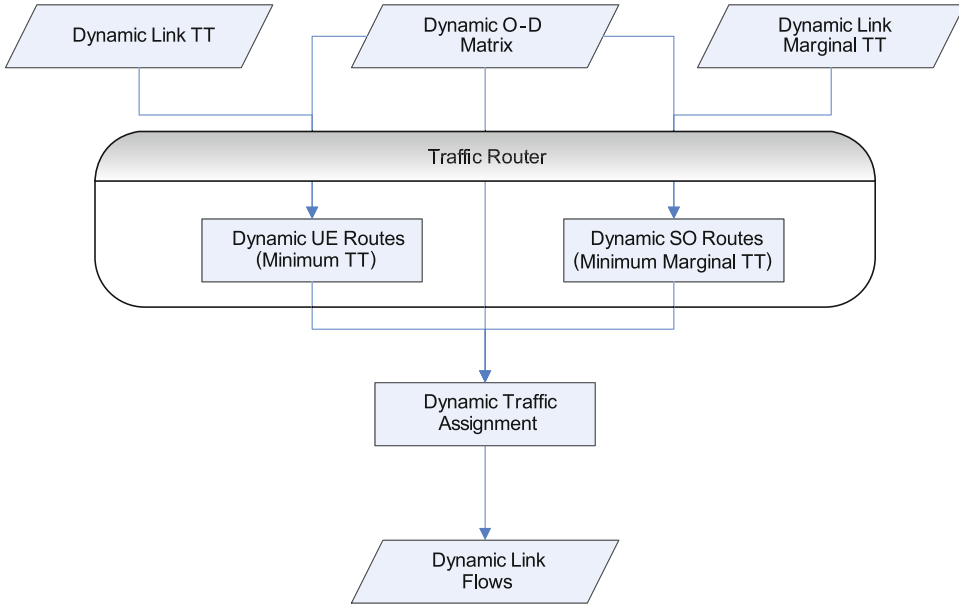
After solving these two formulations the link flows are computed and input into an O-D estimator to provide an estimate of the O-D demand which is then compared to the initial solution. This feedback loop continues until the difference in either link flows or O-D flows is within a desired margin of error or the maximum number of iterations criteria is met.

Mathematical Formulations

Following the notation presented by Sheffi (1985) we present the network notations that are used in the mathematical formulation of a static traffic assignment problem. Initially, the variable definitions are presented followed by the vector definitions (bold variables).

N Set of network nodes

A Set of network arcs (links)



Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks, Fig. 5 Dynamic traffic assignment framework

- R** Set of origin centroids
 S Set of destination centroids
 k_{rs} Set of paths connecting O-D pair $(r-s)$; $r \in R, s \in S$
 x_a Flow on arc (a)
 t_a Travel time on arc (a)
 f_k^{rs} Flow on path (k) connecting O-D pair $(r-s)$
 c_k^{rs} Travel time on path (k) connecting O-D pair $(r-s)$
 q_{rs} Trip rate between origin (r) and destination (s)
 $\delta_{a,k}^{rs}$ Indicator variable:

$$\delta_{a,k}^{rs} = \begin{cases} 1, & \text{if arc } (a) \text{ is on path } (k) \text{ between} \\ & \text{O-D pair } (r-s) \\ 0, & \text{otherwise} \end{cases}$$

Using vector notations (bold variables) the variables are defined as,

- x** Vector of flows on all arcs, $= (\dots, x_a, \dots)$
 t Vector of travel times on all arcs,
 $= (\dots, t_a, \dots)$
 f^{rs} Vector of flows on all paths connecting O-D pair $r-s$, $= (\dots, f_k^{rs}, \dots)$
 f Matrix of flows on all paths connecting all O-D pairs, $= (\dots, f^{rs}, \dots)$

- c^{rs}** Vector of travel times on all paths connecting O-D pair $r-s$, $= (\dots, c_k^{rs}, \dots)$
 c Matrix of travel times on all paths connecting all O-D pairs, $= (\dots, c^{rs}, \dots)$
 q Origin-destination matrix (with elements $= q_{rs}$)
 Δ^{rs} Link-path incidence matrix (with $\delta_{a,k}^{rs}$ elements) for O-D pair $r-s$, as discussed below
 Δ Matrix of link-path incidence matrices (for all O-D pairs), $= (\dots, \Delta^{rs}, \dots)$

The link-path incident matrix is of size equal to the number of links or arcs in the network (number of rows) and number of paths between origin (r) and destination (s) . The element in the a th row, and k th column of Δ^{rs} is $\delta_{a,k}^{rs}$. In other words, $(\Delta^{rs})_{a,k} = \delta_{a,k}^{rs}$.

The following basic relations are fundamental to the mathematical program formulation:

- A link performance function, which is also known as the volume-delay curve or the link congestion function, represents the relationship between flow and travel time on a link (a) ($t_a = t_a(x_a)$).
- The mathematical program formulations assume that travel time on a given link is only dependent

on the flow on the subject link (the model does not capture the effect of opposing flows on the delay of opposed flows), or mathematically

$$\frac{\partial t_a(x_a)}{\partial x_b} = 0 \quad \forall a \neq b \quad \text{and} \quad \frac{\partial t_a(x_a)}{\partial x_a} > 0$$

$\forall a$ where, x_b is the flow on link (b) .

- The travel time on a particular path equals the sum of the travel times on the links comprising that path as $c_k^{rs} = \sum_a t_a \delta_{a,k}^{rs}$ $k \in k_{rs}$, $\forall r \in R$, $\forall s \in S$ or $\mathbf{c} = \mathbf{t} \cdot \mathbf{\Delta}$ considering the vector notation.
- The flow on each link equals the sum of the flows on all paths traversing the subject link as $x_a = \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) \forall a \in A$ or $\mathbf{x} = \mathbf{f} \cdot \mathbf{\Delta}^T$.
- The above formula uses the incidence relationships to express link flows in term of path flows, i.e. $\mathbf{x} = \mathbf{x}(\mathbf{f})$. The incidence relationships also mean that the partial derivative of the link flow can be defined with respect to a particular path flow as follows,

$$\frac{\partial x_a(f)}{\partial f_l^{mn}} = \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) = \delta_{a,l}^{mn},$$

where $\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = 0$ if $k \neq l$ or $r - s \neq m - n$.

Where, f_l^{mn} is the flow on path (l) connecting O-D pair $(m - n)$. Since the function $x_a(f)$ includes a flow summation using the subscripts r , s , and k , the variable with respect to which the derivative is being taken is subscribed by m , n , and l , to avoid the confusion in differentiation.

User Equilibrium

As mentioned earlier, the UE model is based on the assumption that each traveler takes the path that minimizes his/her travel time from their origin to their destination, regardless of any effect this might have on the other network users. In other words, at equilibrium, none of the travelers will be able to reduce their travel times by unilaterally switching to another path. This implies that at equilibrium the link flow pattern is such that the travel times on all of the used paths connecting any given O-D pair will be equal. The travel time on all of these used paths will also be less than or equal to the travel time on any of the unused paths.

The mathematical program that represents this model can be cast using Beckmann's transformation as,

$$\begin{aligned} \min. \quad & z(\mathbf{x}) = \sum_a \int_0^{x_a} t_a(w) dw \\ \text{s.t.} \quad & \sum_k f_k^{rs} = q_{rs} \quad \forall r, s \\ & \quad \quad \quad (\text{Flow conservation constraints}) \\ & f_k^{rs} \geq 0 \quad \forall k, r, s \\ & \quad \quad \quad (\text{Non-negativity constraints}) \\ & x_a = \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) \quad \forall a. \end{aligned} \quad (1)$$

It is worth mentioning that this formulation “has been evident in the transportation literature since the mid-1950's, but its usefulness became apparent only when solution algorithms for this program were developed in the late 1960's and early 1970's” (Sheffi 1985).

In order to prove that the solution of Beckmann's transformation program satisfies the user-equilibrium assignment, first the equivalence conditions will be discussed followed by the uniqueness conditions. In the equivalence conditions it will be shown that the first-order conditions for the minimization program are identical to the equilibrium conditions. Whereas, in the uniqueness conditions, it will be shown that the user-equilibrium equivalent minimization program has only one solution. Hence, proving that the solution of Beckmann's transformation program satisfies the user-equilibrium assignment problem.

Equivalency Conditions Beckmann's transformation program is a minimization program with linear equality and non-negativity constraints. In order to find the first-order conditions for such a program, the Lagrangian with respect to the equality constraints can be written as

$$\begin{aligned} L(f, u) = & z[\mathbf{x}(f)] + \sum_r \\ & \times \sum_s u_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right), \end{aligned} \quad (2)$$

where u_{rs} denotes the dual variable associated with the flow conservation constraint for O-D pair $(r-s)$. At the stationary point of the Lagrangian, the

following first-order conditions have to hold with respect to the path-flow variables and the dual variables. First, with respect to the pathflow variables

$$\begin{aligned} f_k^{rs} \frac{\partial L(f,u)}{\partial f_k^{rs}} &= 0 \quad \forall k,r,s \\ \text{and } \frac{\partial L(f,u)}{\partial f_k^{rs}} &\geq 0 \quad \forall k,r,s \end{aligned} \quad (3)$$

must hold. Alternatively, with respect to the dual variables

$$\frac{\partial L(f,u)}{\partial u_{rs}} = 0 \quad \forall r,s \quad (4)$$

must hold. In addition to the following non-negativity constraints,

$$f_k^{rs} \geq 0 \quad \forall k,r,s. \quad (5)$$

Note that the formulation of this Lagrangian is given in terms of path flow by using the incidence relationships, $x_a = x_a(f)$.

The partial derivative of $L(x, u)$ with respect to the flow variables f_l^{mn} can be given by

$$\begin{aligned} \frac{\partial L(f,y)}{\partial f_l^{mn}} &= \frac{\partial}{\partial f_l^{mn}} z[x(f)] \\ &+ \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s u_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right). \end{aligned} \quad (6)$$

Using the chain rule the first term can be solved as

$$\begin{aligned} \frac{\partial}{\partial f_l^{mn}} z[x(f)] &= \sum_{b \in A} \frac{\partial z(x)}{\partial x_b} \cdot \frac{\partial x_b}{\partial f_l^{mn}} \\ &= \sum_b \left[\left(\frac{\partial}{\partial x_b} \sum_a \int_0^{x_a} t_a(w) dw \right) \cdot \left(\frac{\partial x_b}{\partial f_l^{mn}} \right) \right] \\ &= \sum_b t_b \cdot \delta_{b,l}^{mn} = c_l^{mn}. \end{aligned} \quad (7)$$

The second term can be solved as

$$\frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s u_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right) = -u_{mn}, \quad (8)$$

because (a) u_{rs} is not a function of f_l^{mn} ; (b) q_{rs} is constant; and

$$\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = \begin{cases} 1, & \text{if } r = m, s = n, \text{ and } k = l \\ 0, & \text{otherwise} \end{cases}.$$

Consequently, Eq. (3) and (4) can be solved to derive

$$\frac{\partial L(f,u)}{\partial f_l^{mn}} = c_l^{mn} - u_{mn}.$$

Hence, we can derive the following first-order conditions,

$$\begin{aligned} f_k^{rs} (c_k^{rs} - u_{rs}) &= 0 \quad \forall k,r,s \\ c_k^{rs} - u_{rs} &\geq 0 \quad \forall k,r,s \\ \sum_k f_k^{rs} &= q_{rs} \quad \forall r,s \\ f_k^{rs} &\geq 0 \quad \forall k,r,s. \end{aligned} \quad (9)$$

We can imply the following from these conditions, (1) The first two conditions, for any path (k) connecting any O-D pair ($r - s$), either (a) The flow on that path, f_k^{rs} , equals zero, in which case, the travel time on this path, c_k^{rs} , will have a value that is greater than or equal to the value of the O-D specific Lagrange multiplier, u_{rs} , or, (b) the flow on that path will have a value (greater than zero), in which case, the travel time on this path will have a value equal to the value of the O-D specific Lagrange multiplier, u_{rs} . In both cases, the value of the O-D specific Lagrange multiplier is always less than or equal to the travel time on all other paths connecting the same O-D pair. Hence, this value of the Lagrange multiplier is the minimum path travel time between this O-D pair thus proving that the solution of Beckman's transformation program satisfies the user-equilibrium assignment.

The last two conditions satisfy the flow conservation and non-negativity constraints, respectively. The proof can further be explained as follows, paths connecting O-D pair ($r - s$) can be divided into two groups, (1) Paths with zero flow, and are characterized by a travel time which is either greater than or equal to the minimum travel time; and (2) Paths with non-negative flows, and are characterized by minimum travel times. Thus, confirming the user-equilibrium notion which states that no user can improve his/her travel times by unilaterally changing their routes.

The above proves that user-equilibrium conditions are satisfied at any stationary point of Beckman's transformation program. The following section proves that there is only one solution for Beckman's transformation program. It proves that Beckman's transformation program has only one stationary point, and that this point is a minimum.

Uniqueness Condition In order to prove that Beckmann's transformation program has only one solution, it is sufficient to prove that the objective function is strictly convex in the vicinity of the solution point, convex everywhere else (within the feasible solution region), and that the feasible region (defined by the constraints) is convex.

It is known that linear equality constraints ensure a convex feasible region, and that the addition of the non-negativity constraints does not alter this fact. The convexity of the objective function, with respect to link flows, can be proven in two different ways. The first way can be achieved by the application of the properties of convex functions, on the link-performance functions. On the other hand, the second proof is achieved by proving that the Hessian matrix of the objective function is positive definite.

Link performance functions are known to be continuously increasing functions. Hence, link-performance functions are convex functions. The objective function equals the summation of the integral of the link-performance functions of all links. Properties of convex functions state that integrals of convex functions are also convex functions, and that the summation of convex functions is also a convex function. Hence, proving that the objective function is convex everywhere. Subsequently proving that there is only one solution for Beckman's transformation program, with respect to link flows, and that that solution is a minimum.

Recalling that

$$\frac{\partial^2 z(x)}{\partial x_m \partial x_n} = \frac{\partial t_m(x_m)}{\partial x_n} = \begin{cases} 1, & \text{for } m = n \\ 0, & \text{otherwise} \end{cases}, \quad (10)$$

the Hessian matrix for the objective function can be calculated to be as follows,

$$\begin{aligned} \nabla^2 z(x) &= \begin{bmatrix} \frac{\partial^2 z(x)}{\partial x_1^2} & \frac{\partial^2 z(x)}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 z(x)}{\partial x_A \partial x_1} \\ \frac{\partial^2 z(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 z(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 z(x)}{\partial x_A \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 z(x)}{\partial x_1 \partial x_A} & \frac{\partial^2 z(x)}{\partial x_2 \partial x_A} & \cdots & \frac{\partial^2 z(x)}{\partial x_A^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{dt_1(x_1)}{dx_1} & 0 & \cdots & 0 \\ 0 & \frac{dt_2(x_2)}{dx_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{dt_A(x_A)}{dx_A} \end{bmatrix}. \end{aligned} \quad (11)$$

Obviously, the matrix is definite positive, proving that the objective function is strictly positive, and subsequently, has a unique minimum solution.

It is worth mentioning that the Beckman's transformation program is not convex with respect to path flows, and therefore, the equilibrium conditions themselves are not unique with respect to path flows. In other words, while there is actually only one unique solution for link flows, there are an infinite number of paths flows solutions that would produce this unique link flows solution, which raises the need to compute the most likely of these solutions using a synthetic O-D estimator as was described earlier and will be discussed later in more detail.

System Optimum

As mentioned earlier, the SO model attempts to minimize the total travel time spent in the network. Hence, it might assign certain trips to a slightly longer path (in terms of travel time), in order to reduce the travel time of other user trips by a value which is greater than the value of the increased travel time, and thus achieving a reduced total network travel time. Opposite to user equilibrium, in the system optimum state, users can reduce their travel times by unilaterally switching to alternative paths, which becomes a

challenge to implement such a strategy. Therefore, the solution is not stable. SO network travel time mainly serves as a yardstick that measures the performance of a network.

The mathematical program that represents this model can be written as follows,

$$\begin{aligned}
 \min. & \tilde{z}(x) = \sum_a x_a \cdot t_a(x_a) \\
 \text{s.t.} & \sum_k f_k^{rs} = q_{rs} \quad \forall r,s \\
 & (\text{Flow conservation constraints}) \\
 & f_k^{rs} \geq 0 \quad \forall k,r,s \\
 & (\text{Non-negativity constraints}) \\
 & x_a = \sum_r \sum_s \sum_k \left(f_k^{rs} \cdot \delta_{a,k}^{rs} \right) \quad \forall a.
 \end{aligned} \tag{12}$$

As can be seen, the only difference between user-equilibrium and system optimum programs is the objective function. The SO optimum objective function equals the summation of the products of the travel time on each link times the traffic volume assigned to this link, for all links. Hence, it works on minimizing the total travel time experienced by all vehicles traveling on all links of the networks. On the other hand, the UE objective function equaled the summation of only the travel times of all links.

It can also be seen that the constraints in the SO model are exactly the same as in the UE model. Consequently, similar to the case with the user-equilibrium equivalent program, the solution of this program can be found by solving for the first-order conditions for a stationary point of the following Lagrangian

$$\begin{aligned}
 \tilde{L}(f,u) = & \tilde{z}[x(f)] + \sum_r \\
 & \times \sum_s \tilde{u}_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right), \tag{13}
 \end{aligned}$$

where $\tilde{u}_{rs}$ denotes the dual variable associated with the flow conservation constraint for O-D pair (r,s) . At the stationary point of the Lagrangian, the following first-order conditions have to hold with respect to the path-flow variables

$$\begin{aligned}
 f_k^{rs} \frac{\partial \tilde{L}(f,\tilde{u})}{\partial f_k^{rs}} &= 0 \quad \forall k,r,s \\
 \text{and } \frac{\partial \tilde{L}(f,\tilde{u})}{\partial f_k^{rs}} &\geq 0 \quad \forall k,r,s.
 \end{aligned} \tag{14}$$

With respect to the dual variables

$$\frac{\partial \tilde{L}(f,u)}{\partial \tilde{u}_{rs}} = 0 \quad \forall r,s. \tag{15}$$

In addition to the non-negativity constraints

$$f_k^{rs} \geq 0 \quad \forall k,r,s. \tag{16}$$

Note that, the formulation of this Lagrangian is given in terms of path flow by using the incidence relationships, $x_a = x_a(f)$.

The partial derivative of $\tilde{L}(x,\tilde{u})$ with respect to the flow variables f_l^{mn} can be given by

$$\begin{aligned}
 \frac{\partial \tilde{L}(f,\tilde{u})}{\partial f_l^{mn}} &= \frac{\partial}{\partial f_l^{mn}} \tilde{z}[x(f)] \\
 &+ \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s \tilde{u}_{rs} \left(q_{rs} - \sum_k f_k^{rs} \right) \\
 \frac{\partial}{\partial f_l^{mn}} \tilde{z}[x(f)] &= \sum_{b \in A} \frac{\partial \tilde{z}(x)}{\partial x_b} \cdot \frac{\partial x_b}{\partial f_l^{mn}} \\
 &= \sum_b \left[\left(\frac{\partial}{\partial x_b} \sum_a x_a \cdot t_a(x_a) \right) \frac{\partial x_b}{\partial f_l^{mn}} \right] \\
 &= \sum_b \left[t_b(x_b) + x_b \frac{dt_b(x_b)}{dx_b} \right] \delta_{b,l}^{mn} \\
 &= \sum_b \tilde{t}_b \cdot \delta_{b,l}^{mn} = \tilde{c}_l^{mn}.
 \end{aligned}$$

Assuming $\tilde{t}_b = t_b(x_b) + x_b \frac{dt_b(x_b)}{dx_b}$ and $\partial / (\partial f_l^{mn}) \sum_r \sum_s \tilde{u}_{rs} (q_{rs} - \sum_k f_k^{rs}) = -\tilde{u}_{mn}$ because (a) u_{rs} is not a function of f_l^{mn} ; (b) q_{rs} is constant; and (c)

$$\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = \begin{cases} 1, & \text{if } r = m, s = n, \text{ and } k = l \\ 0, & \text{otherwise.} \end{cases}$$

Therefore $(\partial \tilde{L}(f,\tilde{u}) / (\partial f_l^{mn})) = \tilde{c}_l^{mn} - \tilde{u}_{mn}$.

Where, $\tilde{t}_a$ is a summation of two terms, (1) $t_a(x_a)$, which is the travel time experienced by this

additional driver when the total link flow is (x_a) and (2) $\frac{dt_a(x_a)}{dx_a}$, which is the additional travel time burden that this driver inflicts on each one of the other (x_a) travelers already using link a .

In summary, it can be interpreted as the marginal contribution of an additional traveler – or an infinitesimal flow unit – on the a th link to the total travel time on that link.

Substituting the above results into Eqs. (14) through (16), we get the following first-order conditions

$$\begin{aligned} f_k^{rs} (\tilde{c}_k^{rs} - \tilde{u}_{rs}) &= 0 & \forall k, r, s \\ \tilde{c}_k^{rs} - \tilde{u}_{rs} &\geq 0 & \forall k, r, s \\ \sum_k f_k^{rs} &= q_{rs} & \forall r, s \\ f_k^{rs} &\geq 0 & \forall k, r, s \end{aligned}$$

Similar to the interpretation of the user equilibrium conditions, the following can be implied from the above, (1) the first two Conditions, for any path (k) connecting any O-D pair $(r-s)$, either (a) the flow on that path, f_k^{rs} , equals zero whenever the marginal total travel time on this path, $\tilde{c}_k^{rs}$, will have a value that is greater than or equal to the value of the O-D specific Lagrange multiplier, $\tilde{u}_{rs}$, or, (b) the flow on that path, f_k^{rs} , will have a value (greater than zero) whenever the marginal total travel time on this path, $\tilde{c}_k^{rs}$, will have a value equal to the value of the O-D specific Lagrange multiplier, $\tilde{u}_{rs}$. In both cases, the value of the O-D specific Lagrange multiplier is always less than or equal to the marginal total travel time on all other paths connecting the same O-D pair, i.e. the value of the Lagrange multiplier is the marginal travel time on the used paths between this O-D pair. (2) The last two conditions satisfy the flow conservation and non-negativity constraints, respectively.

The proof can further be explained as follows, paths connecting O-D pair $(r-s)$ can be divided into two groups, (1) Paths with zero flow, and are characterized by a total marginal travel time which is either greater than or equal to the marginal travel time of the used networks (or the Lagrange multiplier). (2) Paths with non-negative

flows, and are characterized by equal marginal travel times.

In order to prove that the SO program has only one solution, as was the case with the user equilibrium program, it is sufficient to prove that the objective function is strictly convex in the vicinity of the solution point, convex everywhere else (within the feasible solution region), and that the feasible region (defined by the constraints) is convex.

It is known that linear equality constraints assure a convex feasible region, and that the addition of the non-negativity constraints does not alter this fact. The convexity of the objective function, with respect to link flows, can be proven if the Hessian matrix of the objective function is positive definite.

Recalling that,

$$\begin{aligned} \frac{\partial^2 \tilde{z}(x)}{\partial x_m \partial x_n} &= \frac{\partial}{\partial x_n} \cdot \left[\frac{\partial}{\partial x_m} \sum_a x_a \cdot t_a(x_a) \right] \\ &= \frac{\partial}{\partial x_n} \left[t_m(x_m) + x_m \frac{dt_m(x_m)}{dx_m} \right] \\ &= \begin{cases} 2 \frac{dt_n(x_n)}{dx_n} + x_n \frac{d^2 t_n(x_n)}{dx_n^2}, & \text{for } m = n \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

As in the user equilibrium program, the Hessian matrix for the objective function can be calculated to be as

$$\begin{aligned} \nabla^2 \tilde{z}(x) &= \begin{bmatrix} 2 \frac{dt_1(x_1)}{dx_1} + x_1 \frac{d^2 t_1(x_1)}{dx_1^2} & 0 & \dots & 0 \\ 0 & 2 \frac{dt_2(x_2)}{dx_2} + x_2 \frac{d^2 t_2(x_2)}{dx_2^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 2 \frac{dt_A(x_A)}{dx_A} + x_A \frac{d^2 t_A(x_A)}{dx_A^2} \end{bmatrix} \end{aligned}$$

This Hessian matrix is positive definite if all the diagonal terms are positive, which is manifested if the link performance functions are positive. Based on the earlier discussion in the

user equilibrium section, it was demonstrated that link-performance functions are convex, and thus demonstrating that the objective function is strictly positive, and subsequently, has a unique minimum solution – with respect to link flows.

It is worth noting that user equilibrium and system optimum produce identical results in any of the following: (1) If congestion effects were ignored, i.e. $t_a(x_a) = t'_a$ (a constant value per arc) or (2) In case of minimal traffic volumes, that would have negligible effects on the arc specific travel times, $t_a(x_a)$.

Dynamic Traffic Assignment Solution Approach

The extension from a static to a dynamic formulation involves the introduction of two time indices into the formulation. The first time index identifies the time at which the path flow leaves its origin while the second time index identifies when the path flow is observed on a specific link. Unfortunately, the introduction of these time indices deems the objective function non-convex and thus two approaches are considered in solving this problem. The first approach is to divide the analysis period into time intervals while assuming that conditions are static within each time interval (time-dependent static or quasi static). The duration of these intervals are network dependent and should be sufficiently long enough to ensure that motorists can complete their trip within the time interval. The static UE and SO mathematical programs can then be solved for each time interval using the standard static formulations that were presented earlier. The mathematical solution approach requires a closed form solution using an analytical modeling approach. Analytical modeling of the network aims at finding the correct mathematical presentation of DTA models that would realistically reflect the real world problem with minimum compromises in the modeling of traffic behavior. The solution of such models should guarantee theoretical existence, uniqueness, and stability. Analytical models are valuable because theoretical insights can be analytically derived. Different analytical network modeling may include mathematical programming

formulations, optimal control formulations, and variational inequality formulations (Peeta and Ziliaskopoulos 2001). Literature within this area of research is extensive. In general, models within the group may be classified into (Peeta and Ziliaskopoulos 2001): i) mathematical programming formulations, as the works of Merchant and Nemhauser (1978a, b), Ho (1980), Carey (1986, 1987, 1992), Janson (1991a, b), Birge and Ho (1993), Ziliaskopoulos (2000), Carey and Subrahmanian (2000); ii) optimal control formulations, as in the works of Friesz et al. (1989), Ran and Shimazaki (1989a, b), Wie (1991), Ran et al. (1993), Boyce et al. (1995); and iii) variational inequality formulations, as with the works of Dafermos (1980), Friesz et al. (1993), Wie et al. (1995), Ran et al. (1996), Ran and Boyce (1996), Chen and Hsueh (1998).

Alternatively, the second approach involves the use of a simulation solution approach. Simulation models on the other hand, in spite of solving the DTA problem within a simulation environment, still use some form of mathematical abstraction of the problem. According to Peeta (Peeta and Ziliaskopoulos 2001), *“the terminology simulation-based models may be a misnomer. This is because the mathematical abstraction of the problem is a typical analytical formulation, mostly of the mathematical programming variety in the current literature. However, the critical constraints that describe the traffic flow propagation, and the spatio-temporal interactions, such as the link-path incidence relationships, flow conservation, and vehicular movements are addressed through simulation instead of analytical evaluation while solving the problem. This is because analytical representations of traffic flows that adequately replicate traffic theoretic relationships and yield well-behaved mathematical formulations are currently unavailable. Hence, the term simulation-based primarily connotes the solution methodology rather than the problem formulation. A key issue with simulation-based models is that theoretical insights cannot be analytically derived as the complex traffic interactions are modeled using simulation. On the other hand, due to the inherently ill-behaved nature of the DTA problem, notions of convergence and*

uniqueness of the associated solution may not be particularly meaningful from a practical standpoint. In addition, due to their better fidelity vis-à-vis realistic traffic modeling, simulation-based models have gained greater acceptability in the context of real-world deployment.”

One of the early simulation DTA tools is the Simulation and Assignment in Urban Road Networks (SATURN) approach. The SATURN algorithm utilizes an equilibrium technique which optimally combines a succession of all-or-nothing assignments (i.e. it is an iterative equilibrium assignment based on iterative traffic loading) (Bolland et al. 1979; Hall et al. 1980; Van Vliet 1982). This model treats platoons of traffic rather than individual vehicles but delays vehicles but delays at intersections are treated in considerable detail. The model consists of two parts: a simulation component and a traffic assignment component. The traffic simulation component fits a delay-flow power curve to three points, namely: zero flow, current flow, and capacity. This delay-flow curve is used by the assignment model to route vehicles. For each traffic signal four cyclic flow profiles are considered: the IN pattern, the ARRIVE pattern, the ACCEPT pattern, and the OUT pattern. SATURN can account for delays caused by opposing flows, delays caused by vehicles on the same roadway, the shape of the arriving platoon, the effect of traffic signal phasing structure and offsets, and individual lane capacities. Arrival rates that exceed capacity are assumed to form queues that build up at constant rates. SATURN can model networks at two levels of detail, namely: inner and buffer. The model was used in studies in the U.K., Australia, and New Zealand. The limitations of the model include: (a) it assumes steady-state conditions for periods of 15–30 min and thus is a time-dependent dynamic assignment approach; (b) queues are modeled vertically and thus they cannot spillback to upstream intersections; (c) it is unsuitable for freeways; (d) it cannot model oversaturated conditions explicitly.

Another early simulation DTA that was developed in the late 1970s is the CONTRAM model (CONtinuous TRaffic Assignment Model). CONTRAM is similar to SATURN in that it

combines traffic assignment with traffic simulation (Leonard et al. 1978). CONTRAM is a computer based time-varying assignment and queuing model. Unlike SATURN, vehicles are grouped within CONTRAM into packets where each packet is treated in the same way as a single vehicle when assigning it to its minimum path. Time varying flow conditions are modeled by dividing the simulation period into a number of consecutive time intervals, which need not be of the same length, and the packets leave each origin at a uniform rate through each such interval. The assignment is an incremental iterative technique where during the first iteration, packets are routed based on link-travel times of previous packets. However, in successive iterations, they are routed based on link travel times that reflect a weighting of travel times during previous iterations and previous packets. Prior to routing a packet, the packet volume is removed from its previously used links. An advantage of this assignment model is that it takes into account the effects of packets leaving later on the routing of packets which leave earlier. Thus, it decides upon the path based on a fully loaded network, rather than on one in which has only been loaded to the extent of any previous increments. This model is more dynamic than most models because vehicles are able to change their routing decisions while en-route, if traffic conditions alter. Satisfactory convergence is usually achieved in 5 to 10 iterations. The limitations of the model are: (a) introduction of signal optimization makes the model unable to converge; (b) vehicles queue vertically on a link; (c) no limitation of the storage capacity of a link is introduced; (d) it is unsuitable for freeway networks; and (e) it can only assign vehicles based on Wardrop's first principle.

A number of contemporary DTA models were developed using the basic CONTRAM concept, including the INTEGRATION (Rakha et al. 1989, 1998; Rilett and Aerde 1993; Rilett and Van Aerde 1991a, b; Rilett et al. 1991; Van Aerde 1985; Van Aerde and Rakha 1989, 2007; Van Aerde and Yagar 1988a, b), DYNASMART (Abdelfatah and Mahmassani 2001; Abdelghany et al. 1999, 2000; Abdelghany and Mahmassani 2001; Chiu and Mahmassani 2001; Jayakrishnan

and Mahmassani 1990, 1991; Jayakrishnan et al. 1993; Peeta et al. 1991; Srinivasan and Mahmassani 2000) and DYNAMIT (Balakrishna et al. 2005; Ben-Akiva et al. 1998a; Koutsopoulos et al. 1995; Yang and Ben-Akiva 2000) modeling approaches. In this section the INTEGRATION dynamic traffic assignment and modeling framework is briefly described as an example illustration of a microscopic traffic assignment and simulation approach. The INTEGRATION model is similar to the CONTRAM model in that it models individual vehicles (packets of unit size). Unlike, other traffic assignment models, the INTEGRATION traffic simulation logic is microscopic in that it models vehicles at a deci-second level of resolution. The software combines car-following, vehicle dynamics, lane-changing, energy, and emission models. Thus, mobile source emissions can be directly estimated from instantaneous speed and acceleration levels. Furthermore, the traffic and emission modeling modules have been tested and validated extensively. For example, the software, which was developed over the past two decades, has not only been validated against standard traffic flow theory (Rakha and Crowther 2002; Rakha and Van Aerde 1996), but has also been utilized for the evaluation of real-life applications (Rakha et al. 2000; Rakha and Zhang 2004b). Furthermore, the INTEGRATION software offers unique capability through the explicit modeling of vehicle dynamics by computing the tractive and resistance forces on the vehicle each deci-second (Rakha et al. 2001a, 2004c; Rakha and Lucic 2002).

The INTEGRATION software uses car-following models to capture the longitudinal interaction of a vehicle and its preceding vehicle in the same lane. The process of car-following is modeled as an equation of motion for steady-state conditions (also referred to as stationary conditions in some literature) plus a number of constraints that govern the behavior of vehicles while moving from one steady-state to another (decelerating and/or accelerating). The first constraint governs the vehicle acceleration behavior, which is typically a function of the vehicle dynamics (Rakha and Crowther 2002; Rakha

et al. 2004b). The second and final constraint ensures that vehicles maintain a safe position relative to the lead vehicle in order to ensure asymptotic stability within the traffic stream. A more detailed description of the longitudinal modeling of vehicle motion is provided by (Rakha et al. 2004d). Alternatively, lane-changing behavior describes the lateral behavior of vehicles along a roadway segment. Lane changing behavior affects the vehicle car-following behavior especially at high intensity lane changing locations such as merge, diverge, and weaving sections.

The INTEGRATION model provides for 7 basic traffic assignment/ routing options: (a) Time-Dependent Method of Successive Averages (MSA); (b) Time-Dependent Sub-Population Feedback Assignment (SFA); (c) Time-Dependent Individual Feedback Assignment (IFA); (d) Time-Dependent Dynamic Traffic Assignment (DTA); (e) Time-Dependent Frank-Wolf Algorithm (FWA); (f) Time-Dependent External; and (g) Distance Based Routing.

The derivation of a time series of MSA traffic assignments involves analyzing each time slice in isolation of either prior or subsequent time slices (time-dependent static or quasi static). The link travel times, upon which the route computations are based, are estimated based on the prevailing O-D pattern and an approximate macroscopic travel time relationship for each link. Multiple paths are computed in an iterative fashion, where the tree for each subsequent iteration is based on the travel times estimated during the previous iterations. The weight assigned to each new tree is $1/N$ where N is the iteration number.

In the case of the feedback assignment vehicles base their routings on the experience of previous vehicle departures (incremental traffic assignment). In the case of the SFA assignment all drivers of a specific type are divided into 5 sub-populations each consisting of 20% of all drivers. The paths for each of these sub-populations are then updated every t seconds during the simulation based on real-time measurements of the link travel times for that specific vehicle class. The value of t is a user-specified value. Furthermore, the minimum path updates of

each vehicle sub-population are staggered in time, in order to avoid having all vehicle sub-populations update their paths at the same time. This results in 20% of the driver paths being updated every $t/5$ s. In the case of the IFA assignment all paths are customized to each individual driver and may therefore be unique relative to any other drivers. This incremental traffic assignment accounts the effect of earlier vehicle departures on the travel time of later which is very similar to the CONTRAM approach.

However, unlike CONTRAM no iterations are made to reassign all the vehicles.

The INTEGRATION DTA computes the minimum path for every scheduled vehicle departure, in view of the link travel times anticipated in the network at the time the vehicle will reach these specific links. The anticipated travel time for each link is estimated based on anticipated link traffic volumes and queue sizes. This routing involves the execution of a complete mesoscopic DTA model prior to the simulation of the traffic. During this DTA, the routes of all vehicles are computed using the above procedure. Upon completion of this DTA, the actual simulation simply implements the routings computed as per the DTA.

Clearly the validity of any of these modeling approaches hinges on the ability of the traffic simulation model to reflect real-life behavior and capture all the complexities of traffic modeling. Clearly, no modeling approach can claim that it is capable of capturing every aspect of empirical traffic flow behavior and thus the output of such models should be interpreted within the context of how they model the spatio-temporal behavior of drivers.

It should also be noted that the models that were described in this section are heuristic approaches attempting to solve the mathematical formulations that were presented earlier and thus there is no guarantee that they converge to a single (unique) solution for UE and/or SO assignment problems for a complex dynamic network. Furthermore, it is not clear if drivers actually attain such an equilibrium state in such networks. Consequently, research is needed to study and develop models on how drivers select routes, how they

respond to the dissemination of traffic information, and how their routing decisions vary temporally in the short- and long-term.

Traffic Modeling

A key component of a DTA is the modeling of traffic stream behavior in order to predict traffic states into the near future and compute link travel times and various measures of effectiveness, as was illustrated earlier in Fig. 2. This section briefly summarizes the various state-of-the-practice approaches to traffic modeling. Researchers have demonstrated that these approaches are unable to predict empirical spatio-temporal aspects (Kerner 2004a) observed in the field. Conversely, others have argued that these approaches, while not perfect, capture the main aspects of empirical data. While our objective is not to argue either way, it is sufficient to note that these tools are being used by transportation professionals to assess dynamic networks and thus are presented in this section. These approaches can be classified into three categories, which include: macroscopic, mesoscopic, and microscopic approaches. Each of these approaches is briefly described in this section. Again the description is by no means comprehensive but does provide a general overview of these approaches. The interested reader should consider reading the wealth of literature on this topic.

Prior to describing the specifics of the various modeling approaches it is important to note that with the exception to research conducted by Kerner (2004a, b), most existing approaches are based on the famous one-dimensional kinematic waves (KW) theory, which was proposed by Lighthill and Witham (1955) and independently proposed by Richards (1956). The key postulate of the theory is that there exists a functional relationship between the traffic stream flow rate q and density k that might vary with location x but does not vary with time t (this contradicts the definition of dynamic given that within a dynamic process variables vary spatially and temporarily). It should be noted that in microscopic approaches, as will be described later, the fundamental

diagram various temporally as a function of the traffic composition, thus overcoming some of the drawbacks of this approach. The fundamental hypothesis of all traffic flow theories is the existence of a site-specific unique relationship between traffic stream flow and traffic stream density, commonly known as the fundamental diagram, the traffic stream motion model, or the car-following model at the microscopic level. The assumption is that all steady-state model solutions lie on the fundamental diagram and thus are referred to as fundamental diagram approaches (Kerner 2004a). Given that traffic stream space-mean speed can be related to traffic stream flow and density, a unique speed-flow-density relationship (in the macroscopic approach) is derived from the fundamental diagram for each roadway segment. This relationship can also be cast at the micro-level by relating the vehicle speed to its spacing, given that vehicle spacing is the inverse of traffic stream density. Some researchers have argued that the fundamental diagram approach cannot capture the spontaneous traffic stream failure that is observed in the field and thus these researchers have proposed other theories.

One of these theories is the three-phase traffic flow theory proposed by Kerner (2004a, b), which attempts to explain empirical spatiotemporal features of congested patterns. The theory divides traffic into three phases: free-flow, synchronized flow, and wide moving jams. The free-flow phase is consistent with the uncongested regime on a fundamental diagram and thus is not discussed further. The *synchronized flow* phase involves continuous traffic flow with no significant stoppage. The word “flow” reflects this feature. Within this phase there is a tendency towards synchronization of vehicle speeds and flows across the different lanes on a multilane roadway, and thus comes the name “synchronized.” This synchronization of speeds is a result if the relatively low probability of passing within this phase. The third phase, *wide moving jam*, is a phase that involves traffic jams that propagate through other states of traffic flow and through any bottleneck while maintaining the velocity of the downstream jam front. The phrase “moving jam” reflects the propagation as a whole localized structure on a road.

To distinguish wide moving jams from other moving jams, which do not characteristically maintain the mean velocity of the downstream jam front, Kerner uses the term wide moving jam. Kerner indicates that if a moving jam has a width (in the longitudinal direction) considerably greater than the widths of the jam fronts, and if vehicle speeds inside the jam are zero, the jam always exhibits the characteristic feature of maintaining the velocity of the downstream jam front.

Kerner distinguishes his three-phase traffic flow theory from fundamental diagram approaches in a number of aspects. He demonstrates that the fundamental diagram approach cannot capture two key empirically observed phenomena in traffic, namely: (a) the probabilistic nature of free-flow to synchronized flow transition (flow breakdown), and (b) the spontaneous formation of general patterns (GP), which include moving and wide moving jams. Alternatively, it could be hypothesized that by modeling individual driver behavior (micro or nano modeling), capturing vehicle acceleration constraints, and introducing stochastic differences between drivers that this may be sufficient to model these two key phenomena.

Macroscopic Modeling Approaches

In order to solve for the three traffic stream variables (q , k , and u) three equations are introduced. The first is the functional relationship between flow and density, or what is commonly known as the fundamental diagram. Typical functions include the Pipes triangular function (3 parameters), the Greenshields parabolic function (Greenshields 1934) (2 parameters), or the Van Aerde (Rakha and Lucic 2002) function (4 parameters). The second equation is the flow conservation equation (equation of continuity) that can be expressed as $\partial k(x, t) / \partial t + \partial q(x, t) / \partial x = 0$, considering no entering or exiting traffic. The third and final equation relates the traffic stream flow rate (q) to the traffic stream density (k) and space-mean speed (u) as $q = ku$. The numerical solution of the KW problem involves partitioning the network into small cells of length Δx and discretizing time into steps of duration Δt . For numerical stability $\Delta x = u \Delta t$. The problem is solved by

stepping through time and solving for the variables in every cell using the incremental transfer (IT) principle (essentially explicit finite difference method). Extensions to the standard KW solution have introduced IT solutions for each lane along a freeway where the freeway is modeled as a set of interacting streams linked by lane changes. Lane-changing vehicles can be treated as a fluid that can accelerate instantaneously, however this approach does not capture the reduction in capacity that is associated with lane changes. Consequently, further improvements have been introduced through the use of a hybrid approach (Laval and Daganzo 2006) that combines microscopic and macroscopic models. Specifically, slow vehicles are treated as moving bottlenecks in a single KW stream, while lane changing vehicles are modeled as discrete particles with constrained motion. The model requires identifying a lane-change intensity parameter in addition to the functional relationship parameters that were described earlier. It is not clear how such a parameter is derived. The major drawbacks of this modeling approach are that it does not account for the dynamic changes in roadway capacity (e.g. the capacity of a weaving section varies as a function of the traffic composition), it cannot capture the spontaneous traffic stream failure that is observed in the field, it cannot capture the impact of opposing flows on the traffic behavior of an opposed flow (e.g. how the capacity of an opposed left turn movement is affected by the opposing through movement), and it ignores the stochastic nature of traffic.

Mesoscopic Modeling Approaches

The mesoscopic analysis tracks individual vehicles as they travel through the network along a sequence of links that are determined by the traffic assignment. The level of tracking involves computing the vehicle's travel speed on each link based upon the density on the link together with a user specified speed/density relationship. The vehicle is then held on the link for the duration of its travel time. At the vehicle's scheduled departure time, the vehicle is allowed to exit the link if the link privileges permit it to leave; otherwise, the vehicle is held on the link until the link privileges so permit. Link exit privileges may be

controlled by traffic signals at the downstream end of the link or by any queues that may be present on the lane. Queues are stored for each lane separately to account for any queue length differentials that may occur (e.g., longer queues on left turn opposed lanes). The mesoscopic analysis captures the operational level of detail (e.g., the reduction in lane capacity as a result of an opposed flow) without having to track each vehicle's instantaneous speed profile. This means that the computational requirements for such a type of modeling are more than that required by a macroscopic analysis, but less than that required by a microscopic analysis. The INTEGRATION 1.50, DynaSMART, and DynaMIT models are examples of such modeling approaches. This approach suffers from similar drawbacks as identified in the macroscopic analysis procedures, namely an inability to capture correct spatiotemporal propagation of congestion, a failure to capture dynamic changes in capacity, a failure to capture for spontaneous breakdown in a traffic stream, and failure to capture the stochastic nature of traffic.

Microscopic Modeling Approaches

The third approach to modeling traffic is the microscopic analysis, which tracks each vehicle as it travels through the network on a second-by-second or deci-second level of resolution using detailed car-following and lane-changing models. Microscopic simulation software use car-following models to capture the longitudinal interaction of a vehicle and its preceding vehicle in the same lane. The process of car-following is modeled as an equation of motion for steady-state conditions (also referred to as stationary conditions in some literature) plus a number of constraints that govern the behavior of vehicles while moving from one steady-state to another (decelerating and/or accelerating). The first constraint governs the vehicle acceleration behavior, which is typically a function of the vehicle dynamics. The second and final constraint ensures that vehicles maintain a safe position relative to the lead vehicle in order to ensure asymptotic stability within the traffic stream. A more detailed description of the longitudinal modeling of vehicle motion is provided by Rakha et al. (2004d).

While there are a number of commercially available software packages that simulate traffic microscopically (CORSIM, Paramics, FREEVU, VISSIM, AIMSUN2, and INTEGRATION), these approaches are computationally intensive and cannot run in real-time. The INTEGRATION software has been able to capture the stochastic nature of traffic stream capacity by randomly modeling vehicle-specific car-following models. Furthermore, the model captures the capacity loss associated with recovery from breakdown through the vehicle acceleration constraints. The stochastic nature of car-following and lane-changing behavior may allow the model to capture spontaneous breakdown in traffic stream flow.

The amount of computation and memory necessary for simulating a large transportation network at a level of detail down to an individual traveler and an individual vehicle may be extensive. Hence a microscopic massively parallel simulation approach entitled “cellular automata” (CA) is sometimes proposed to simulate large networks. The cellular automata approach essentially divides every link on the network into a finite number of cells. At a one second time step, each of these “cells” is scanned for a vehicle presence. If a vehicle is present, the vehicle position is advanced, either within the cell or to another cell, using a simple rule set (Nagel 1996; Nagel and Schreckenberg 1992, 1995). The rule set is made simple to increase the computational speed necessary for a large simulation. Vehicles are moved from one grid cell to another based on the available gaps ahead, with modifications to support lane changing and plan following, until they reach the end of the grid. There, they wait for an acceptable gap in the traffic or for protection at a signal before moving through the intersection onto the next grid. This continues until each vehicle reaches its destination, where it is removed from the grid. Reducing the size of the “cell,” expanding the rule set, and adding vehicle attributes increases the fidelity of the simulator, but also greatly affects the computational speed. The size of 7.5 m in length and a traffic lane in width is often chosen as a default size for the “cell” as was applied with the TRANSIMS software (Nagel and

Schreckenberg 1992). This approach suffers from a number of drawbacks including the inability to capture the dynamic nature of roadway capacity, the inability to capture spontaneous breakdown of traffic stream, and the inability to capture opposing flow impacts on opposed flow saturation flow rates (e.g. the impact an opposing through movement flow has on the capacity of a permitted left turn movement that has to find a gap in this opposing flow).

Dynamic Travel Time Estimation

As was demonstrated earlier in the paper, the DTA requires arc (link) travel times in order to compute minimum paths. There are several systems commercially available that are capable of estimating real-time travel times. These can be broadly classified into spot speed measurement systems, spatial travel time systems, and probe vehicle technologies. Spot speed measurement systems, specifically inductance loop detectors, have been the main source of real-time traffic information for the past two decades. Other technologies for measuring spot speeds have also evolved, such as infrared and radar technologies. Regardless of the technology, the spot measurement approaches only measure traffic stream speeds over a short roadway segment at fixed locations along a roadway. These spot speed measurements are used to compute spatial travel times over an entire trip using space-mean-speed estimates. In addition, new approaches that match vehicles based on their lengths have also been developed (Coifman 1998; Coifman and Banerjee 2002; Coifman and Cassidy 2001; Coifman and Ergueta 2003). However, these approaches require raw loop detector data as opposed to typical 20- or 30-s aggregated data. Alternatively, spatial travel time measurement systems use fixed location equipment to identify and track a subset of vehicles in the traffic stream. By matching the unique vehicle identifications at different reader locations, spatial estimates of travel times can be computed. Typical technologies include AVI and license-plate video detection systems. Finally, probe vehicle technologies track a sample of probe vehicles on a

second-by-second basis as they travel within a transportation network. These emerging technologies include cellular geo-location, Global Positioning Systems (GPS), and Automatic Vehicle Location (AVL) systems.

Traffic routing strategies under recurring and non-recurring strategies should be based on forecasting of future traffic conditions rather than historical and/or current conditions. In general the traffic prediction approaches can be categorized into three broad areas: (i) statistical models, (ii) macroscopic models, and (iii) route choice models based on dynamic traffic assignment (Ben-Akiva et al. 1992, 1997, 1998b; Birge and Ho 1993; Peeta and Mahmassani 1995a). Time series models have been used in traffic forecasting mainly because of their strong potential for online implementation. Early examples of such approaches include (Ahmed and Cook 1982) and more recently (Ishak and Al-Deek 2003; Lee and Fambro 1999). In addition, researchers have applied Artificial Neural Network (ANN) techniques for the prediction of roadway travel times (Abdulhai et al. 2002; Park 2002; Park and Rilett 1998, 1999; Park et al. 1998; Pavlis and Papageorgiou 1999). These studies demonstrated that prediction errors were affected by a number of variables pertinent to traffic flow prediction such as spatial coverage of surveillance instrumentation, the extent of the loop-back interval, data resolution, and data accuracy.

An earlier publication (Dion and Rakha 2006) developed a low-pass adaptive filtering algorithm for predicting average roadway travel times using Automatic Vehicle Identification (AVI) data. The algorithm is unique in three aspects. First, it is designed to handle both stable (constant mean) and unstable (varying mean) traffic conditions. Second, the algorithm can be successfully applied for low levels of market penetration (less than 1%). Third, the algorithm works for both freeway and signalized arterial roadways. The proposed algorithm utilizes a robust data-filtering procedure that identifies valid data within a dynamically varying validity window. The size of the validity window varies as a function of the number of observations within the current sampling interval, the number of observations in the previous

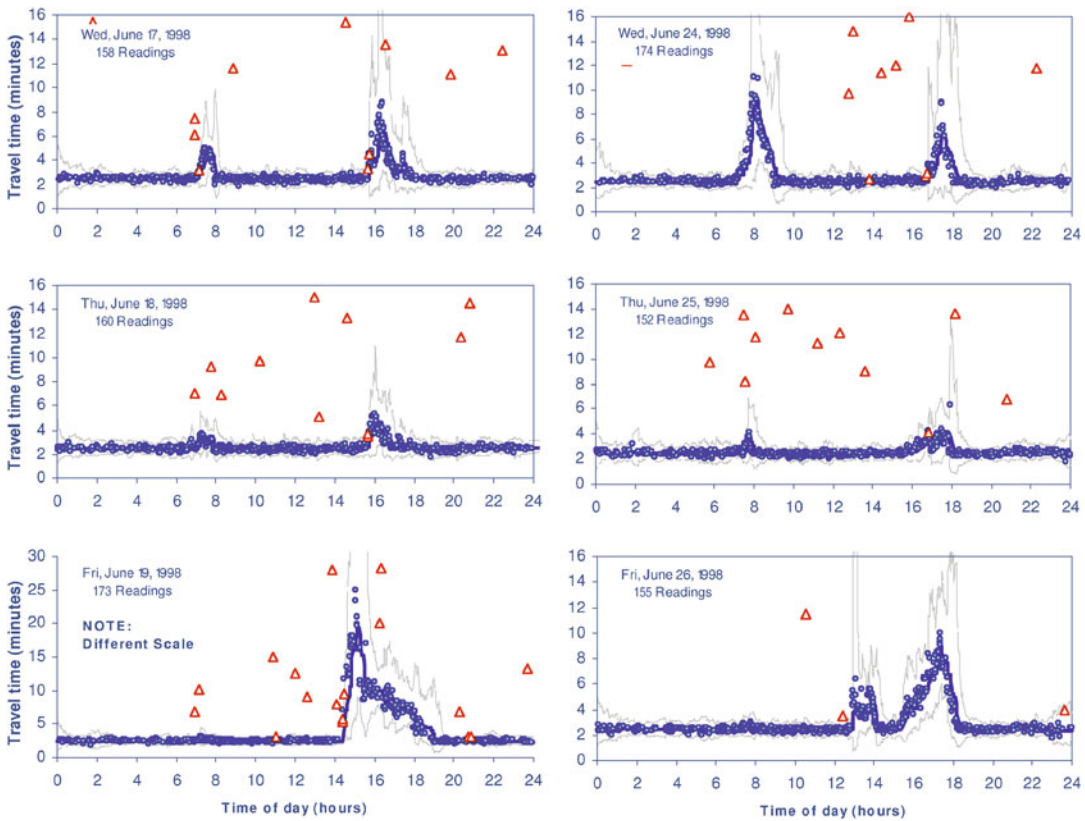
intervals, and the number of consecutive observations outside the validity window. Applications of the algorithm to two AVI datasets from San Antonio, one from a freeway link and the other from an arterial link, demonstrated the ability of the proposed algorithm to efficiently track typical variations in average link travel times while suppressing high frequency noise signals.

Within the filtering algorithm, the expected average travel time and travel time variance for a given sampling interval are computed using a moving average (MA) technique. As shown in Eqs. (17) and (18), it estimates the expected average travel time and expected travel time variance within a given sampling interval based on the set of valid travel time observations in the previous sampling interval and the corresponding previously smoothed moving average value using an adaptive exponential smoothing technique. In both equations, calculations of the smoothed average travel time and travel time variance are made using a lognormal distribution to reflect the fact that the distribution is right skewed (skewed towards longer travel times). Field data from the San Antonio AVI system demonstrated that this assumption is reasonable.

$$\tilde{t}_{i,k} = \begin{cases} e^{\left[\alpha \cdot \ln t_{i,k-1} + (1-\alpha) \cdot \ln \tilde{t}_{i,k-1}\right]}, & \text{if } n_{i,k} > 0, \\ \tilde{t}_{i,k-1}, & \text{if } n_{i,k} = 0, \end{cases} \quad (17)$$

$$\sigma_{i,k}^2 = \begin{cases} \alpha \cdot \sigma_{i,k-1}^2 + (1-\alpha) \cdot \tilde{\sigma}_{i,k-1}^2, & \text{if } n_{i,k} > 1, \\ \tilde{\sigma}_{i,k-1}^2, & \text{if } n_{i,k} \leq 1. \end{cases} \quad (18)$$

It should be noted that $t_{i,k}$ is the observed average travel time along link i within the k th sampling interval (s), $\tilde{t}_{i,k}$ is the smoothed average travel time along link i in the k th sampling interval (s), $\sigma_{i,k}^2$ is the variance of the observed travel times relative to the observed average travel time in the k th sampling interval (s^2), $\tilde{\sigma}_{i,k}^2$ is the variance of the observed travel times relative to the smoothed travel time in the k th sampling interval (s^2), $n_{i,k}$ is the number of valid travel time readings on link i in the k th sampling interval, and $\alpha = 1 - (1 - \beta)^{n_{i,k}}$



Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks, Fig. 6 Sample application of AVI travel time estimation algorithm (Dion and Rakha 2006)

for all i and k is an exponential smoothing factor that varies as a function of the number of observations $n_{i,k}$ within the sampling interval, where β is a constant that varies between 0 and 1.

Figure 6 shows an example application of the algorithm using AVI data along I-35 in San Antonio, TX. The figure illustrates the average travel time estimate (thick line), the validity window bounds (thin lines), what are considered to be valid data (circular), and the observations that are considered to be outliers (triangles). The figure clearly illustrates the effectiveness of the algorithm in estimating roadway travel times for low levels of market penetration of AVI tags.

Once link travel times have been estimated, the expected trip or path travel times can be computed by summing the relevant smoothed link travel times. In addition, the trip travel time reliability, which is the probability that a trip can reach its destination

within a given period at a given time of day, can be computed for use in traffic routing. Travel time reliability is a measure of the stability of travel time, and therefore is subject to fluctuations in flow (Bell and Iida 1997). Typically, when flow fluctuations are large, travel time is often longer than expected. As levels of congestion in transportation networks grow, generally the stability of travel time will have greater significance to transportation users. The trip travel time reliability can be computed as the probability $P(T \leq t)$ that the trip travel time (T) is less than some arbitrary travel time (t), using the cumulative distribution function estimated from an analysis of AVI field data. The current state-of-the-art in predicting trip travel time variability is to assume that the travel times on all the links along a path are generated by statistically independent normal distributions. Consequently, the trip variance can be computed as the summation of the link travel

time variances for all links along a path. As part of the proposed research effort, different statistical techniques (not assuming independent normal variates) will be devised to estimate the trip travel time variance, as discussed in the Proposed Research Tasks section. These techniques will be tested using data from the video detection system that is currently implemented in the Blacksburg Area.

In addition, research has been conducted to estimate the optimum locations of surveillance equipment for the estimation of travel times. Specifically, an earlier publication developed an algorithm for optimally locating Automatic Vehicle Identification tag readers by maximizing the benefit that would accrue from measuring travel times on a transportation network (Sherali et al. 2006). The problem is formulated as a quadratic 0–1 optimization problem where the objective function parameters represent benefit factors that capture the relevance of measuring travel times as reflected by the demand and travel time variability along specified trips. An optimization approach based on the Reformulation-Linearization Technique coupled with semi-definite programming concepts was designed to solve the formulated reader location problem. Alternatively, a Genetic Algorithm (GA) approach was developed to optimally locate the AVI readers (Arafeh and Rakha 2005).

Dynamic or Time-Dependent Origin-Destination Estimation

As was demonstrated earlier the Bechmann UE and the SO formulations do not provide unique path flows and thus a synthetic O-D estimator is required to estimate the path flows from the unique link flows. The techniques used to estimate O-D demands can be categorized based on different factors, as will be discussed in detail. The first categorization, of the available O-D estimation techniques, relates to whether the O-D's to be estimated are static, and apply to only one observation time period, or whether estimates are required for a series of linked dynamic time periods. The next breakdown relates to whether the estimation is based on information about the magnitude of trip ends only, or whether

information is available on additional links along the route of each trip. The former problem is commonly referred to as the trip distribution problem in demand forecasting, while the latter problem is commonly referred to as the synthetic O-D generation problem. Both problems are discussed, but this section will focus on the latter synthetic O-D generation problem. The former is viewed simply as a simpler subset of the latter.

Within the overall static synthetic O-D generation problem, there are two main flavors. The first exists when the routes that vehicles take through the network are known a priori. The second arises when these routes need to be estimated concurrently while the O-D is being estimated. A priori knowledge of routes can arise automatically when there is only one feasible route between each O-D pair, or when observed traffic volumes are only provided for the zone connectors at the origins and destinations in the network. The first condition is common when O-D's are estimated for a single intersection or arterial, or a single interchange or freeway. The second condition is the default for any trip distribution analysis. This section will initially focus on situations where the routes are known a priori. Subsequently, a solution to the more general problem which involves an iterative use of the solution approach when routes are not known a priori will be discussed.

Within the static/dynamic synthetic O-D generation problem, for scenarios where routes are known a priori (or are assumed to be known a priori) there exist two sub-problems. The first of these problems relates to situations where flow continuity exists at each node in the network, and multiple O-D matrices can be shown to match these observed flows exactly. In this case, the most likely of these multiple O-D matrices needs to be identified. The second sub-problem relates to situations where flow continuity does not exist at either the node level or at the network level. In other words, the observed traffic flows are such that no matrix exists that will match the observed flows exactly. In this case, Van Aerde et al. (2003) introduced a new set of complementary link flows that maintain flow continuity by introducing minimum alterations to the observed flows to solve the maximum likelihood problem.

The static synthetic O-D generation problem, for scenarios where flow continuity does exit, can be formulated in two different ways (Van Zuylen and Willumsen 1980; Willumsen 1978). The first of these considers that the fundamental unit of measure is the individual trip, while the second considers that the fundamental unit of measure is the observation of a single vehicle on a particular link. The availability of a seed or target O-D matrix is implicit in the latter formulation, but can be dropped in the former formulation, as was demonstrated in an earlier publication (Van Aerde et al. 2003). However, only when a seed matrix is properly included in the former formulation is it guaranteed to yield consistent results with the latter formulation. In other words, the absence of a seed matrix in the trip based formulation can be shown to yield inconsistent results, at least for some networks in which the multiple solutions result in a different number of total trips.

An additional and related attribute, of the trip-based formulation of maximum likelihood, is the presence of a term in the objective function that is based on the total number of trips in the network. This term, referred to as T , is often dropped in some approximations. However, it can be shown that dropping this term can yield solutions that represent only a very poor approximation to the true solution (Rakha et al. 2005a). In contrast, approximations involve the use of Stirling's approximation, for representing the logarithm of factorials, were shown to yield consistently very good approximations (Rakha et al. 2005a). This finding is critical because use of Stirling's approximation is critical to being able to compute the derivatives that are needed to numerically solve the problem (it is difficult to take derivatives of terms that include factorials).

Other examples from literature include the works of Cremer and Keller (1987), Cascetta et al. (Cascetta and Marquis 1993), Wu and Chang (1996), Sherali et al. (1997), Ashok and Ben-Akiva (2000), Hu et al. (2001), Chang and Tao (1999), Pavlis and Papageorgiou (1999), Peeta and Zhou (1999a, b), Peeta and Yang (2000, 2003), Yang (2001), Peeta and Bulusu (1999).

Comparison of Synthetic O-D and Trip Distribution Formulations

Within the four-step planning process O-D matrices are estimated in the trip distribution step. Several methods are used for trip distribution including the gravity, growth factor, and intervening opportunities models. The gravity model is most utilized because it uses the attributes of the transportation system and land-use characteristics and has been calibrated and applied extensively to the modeling of numerous urban areas. The model assumes that the number of trips between two zones i and j (T_{ij}) is directly proportional to the number of trip productions from the origin zone (P_i) and the number of attractions to the destination zone (A_j) and inversely proportional to a function of travel time between the two zones (F_{ij}) as

$$T_{ij} = P_i \left(\frac{A_j F_{ij} K_{ij}}{\sum_j A_j F_{ij} K_{ij}} \right). \quad (19)$$

Typically the values of trip productions and attractions are computed based on trip generation procedures. The values of F_{ij} are computed using a calibration procedure that involves matching modeled and field trip length distributions. The socio-economic adjustment factors (K_{ij}) values are used when the estimated trip interchange must be adjusted to ensure that it agrees with observed trips by attempting to account for factors other than travel time. The values of K are determined in the calibration process, but considered judiciously when a zone is considered to possess unique characteristics.

Because the O-D problem is under-specified, multiple O-D demands can generate identical link flows. For example, if one attempts to estimate an O-D matrix for a 100 zone network with, say 1000 links, one has more unknowns to solve for than there are constraints. In the case of the trip distribution process, there are 100×100 O-D cells to estimate, and only 2×100 trip end constraints. In the case of the synthetic O-D generation process, there are again 100×100 O-D cells to estimate, and only 1000 link constraints. Given the

possibility of multiple solutions, both the trip distribution process and the synthetic O-D generation process invoke additional considerations to select a preferred matrix from among the multiple solutions.

In the case of synthetic O-D generation, the desire is to select from among all of the possible solutions, the most likely. This approach requires one to define a measure of the likelihood of each matrix. In general, there are two approaches to establish the likelihood of a matrix. One of them treats the trip as the basic unit of observation, while the other considers a volume count as the basic unit of observation. The first approach will be discussed in detail, while the interested reader might refer to the literature (Van Aerde et al. 2003) for a more detailed description of the various formulations. It suffices to indicate that for any matrix with cells T_{ij} , the likelihood of the matrix can be estimated using a function $L = f(T_{ij}, t_{ij})$, where t_{ij} represents prior information. The prior information is often referred to as a seed matrix, and can be derived from a previous study or survey. In the absence of such prior information, all of the cells in this prior matrix should be set to a uniform set of values.

In the case of the trip distribution process, the additional information that is added is some form of impedance. For example, the original gravity model considered that the likelihood of trips between two zones was proportional to the inverse of the square of the distance between the two zones. Since that time, many more sophisticated forms of impedance have been considered, but for the purposes of this discussion, all of these variations can be generalized as being of the form F_{ij} , where $F_{ij} = f(c_{ij})$ or the generalized cost of inter-zonal travel. What is less obvious, however, is the fact that the use of this set of impedance factors F_{ij} , is essentially equivalent to the use of a seed matrix t_{ij} .

Van Aerde et al. (2003) demonstrated that solving the trip distribution problem, using zonal trip productions and attractions as constraints, together with a trip impedance matrix, is essentially the same as solving the synthetic O-D problem using zone connector in and out flows as constraints, and utilizing a seed matrix based on Eq. (20).

$$t_{ij} = T \frac{F_{ij}}{\sum_{ij} F_{ij}} \quad (20)$$

Static Formulations

Entropy maximization and information minimization techniques have been used to solve a number of transportation problems (Wilson 1970). The application of the entropy maximization principles to the static O-D estimation problem was first introduced by Willumsen (1978; Van Zuylen and Willumsen 1980). Willumsen demonstrated that by maximizing the entropy, the most likely trip matrix could be estimated subject to a set of constraints.

The trip-based approach to defining maximum likelihood considers that the overall trip matrix is made up of uniquely identifiable individual trip makers. The most likely matrix is one where the likelihood function is maximized as

$$\max. \quad Z_1(T_{ij}) = \frac{T!}{\prod_{ij} (T_{ij}!)} \quad (21)$$

The above formulation does not take into account any prior information, from for example a previous survey. While the seed matrix does not necessarily have to satisfy the observed link flows, the seed matrix can be utilized to expand the maximum likelihood function to

$$\max. \quad Z_2(T_{ij}, t_{ij}) = \frac{T!}{\prod_{ij} (T_{ij}!)} \prod_{ij} \left(\frac{t_{ij}}{\sum_{ij} t_{ij}} \right)^{T_{ij}} \quad (22)$$

It can be noted that the likelihood of an individual trip from i to j is $t_{ij} / \sum \sum t_{ij}$, based on the above seed matrix. Consequently, the probability of T_{ij} trips being drawn is $(t_{ij} / \sum \sum t_{ij})^{T_{ij}}$.

The above formulations of objective functions for expressing likelihood require additional constraints in order to be complete (Van Zuylen and Willumsen 1980; Willumsen 1978). The simplest of these constraints indicate that the sum of all trips crossing a given link must be equal to the link flow on that link as

$$V_a = \sum_{ij} T_{ij} p_{ij}^a \quad \forall a. \quad (23)$$

As will be shown later, the simplest mechanism, for including the above constraints in the earlier objective functions, is to utilize Lagrange multipliers. These multipliers permit an objective function with equality constraints to be transformed into an equivalent unconstrained objective function.

This simple set of equality constraints, while making the formulation complete, may at times also render the problem infeasible. A more general formulation that was proposed in the literature (Van Aerde et al. 2003) is to minimize the link flow error, rather than eliminate the error. In other words, rather than finding the most likely O-D that exactly replicates the observed link flows, the problem is re-formulated as finding the most likely O-D matrix from among all of those that come equally close to matching the link flows. One expression that is proposed to capture the error to be minimized is shown in Eq. (24), and is subject to the flow continuity constraints in Eq. (25). The constraints in Eq. (25) can be introduced in Eq. (25) to yield an unconstrained objective function, yielding a set of complementary link flows V'_a . These complementary flows are those which deviate the least from the observed link flows, while satisfying link flow continuity. Given that these complementary link flows do satisfy flow continuity, they can now be added in as rigid equality constraints to the objective function (21) or (22), and be guaranteed to yield a feasible solution.

$$\min. \quad Z_3(T_{ij}) = \sum_a (V_a - V'_a)^2 \quad (24)$$

$$V'_a = \sum_{ij} T_{ij} p_{ij}^a \quad \forall a. \quad (25)$$

Alternatively, one can incorporate Eq. (25) into Eq. (24) to yield

$$\min. \quad Z_4(T_{ij}) = \sum_a \left(V_a - \sum_{ij} T_{ij} p_{ij}^a \right)^2 \quad (26)$$

This equation should be minimized concurrently to maximizing the objective function (21) or (22). Unfortunately, it is not easy to combine one expression that desires to maximize

likelihood with another that desires to minimize link flow error, as a Lagrangian can only add equality constraints to a constrained objective function. Van Aerde et al. proposed a solution to this problem which involves taking the partial derivatives of Eq. (26) with respect to each of the trip cells that are to be estimated as

$$\frac{\partial}{\partial T_{ij}} Z_4(T_{ij}) = \frac{\partial}{\partial T_{ij}} \sum_a \left(V_a - \sum_{ij} T_{ij} p_{ij}^a \right)^2 = 0 \quad \forall i,j. \quad (27)$$

$$0 = 2 \left(\sum_a (V_a \cdot p_{ij}^a) - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \quad \forall i,j \quad (28)$$

This yields as many equations as there are trip cells, as shown in Eq. (27). Furthermore, setting these derivatives equal to 0 is equivalent to minimizing Eq. (27). However, while Eq. (24) could not be added to the maximum likelihood objective function, the equalities in Eq. (28) can. This produces an unconstrained objective function that always yields a feasible solution computed as

$$\max. \quad \frac{T!}{\prod_{ij} T_{ij}} \prod_{ij} \binom{t_{ij}}{t}^{T_{ij}} - \sum_{ij} \left(\lambda_{ij} \cdot 2 \left(\sum_a (V_a \cdot p_{ij}^a) - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \right), \quad (29)$$

where: $T = \sum_{ij} T_{ij}$ and $t = \sum_{ij} t_{ij}$.

The net result, of the above process, is to suggest that most synthetic O-D generation problems consist of two sub-problems. One of these involves finding a new set of complementary link flows that do produce link flow continuity, at which point the maximum likelihood problem can be solved as before. Alternatively, one can compute the partial derivatives, that will yield link flow continuity, while deviating by the least amount from the observed link flows, and then utilize them directly in the maximum likelihood formulation using Lagrange multipliers. Both solutions can be shown to yield identical results.

A first challenge with maximizing Eq. (29) is that it yields very large numbers that are difficult to work with. Furthermore, as it is common to maximize objective functions by taking their derivatives, and as it is more difficult to contemplate the derivative of a discontinuous expression, such as those including factorials, a simple approximation is made. This approximation involves taking the natural logarithm of either objective function Eq. (21) or (22). Taking the natural logarithm of the objective function both makes the output easier to handle and permits the use of Stirling's approximation as a convenient continuous equivalent to the term $\ln(x!)$ as

$$\ln(T!) = T \ln T - T \quad (30)$$

The resulting converted objective function using the Stirling approximation on the original objective function of Eq. (22) is computed as

$$\begin{aligned} \max. \quad & T \ln \left(\frac{T}{t} \right) - T - \sum_{ij} \left(T_{ij} \ln \left(\frac{T_{ij}}{t_{ij}} \right) - T_{ij} \right) \\ & - \sum_{ij} \left(\lambda_{ij} \cdot 2 \left(\sum_a \left(V_a \cdot p_{ij}^a \right) - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \right). \end{aligned} \quad (31)$$

Expanding and simplifying the various terms we derive

$$\begin{aligned} & T \ln \left(\frac{T}{t} \right) - T - \sum_{ij} \left(T_{ij} \ln \left(\frac{T_{ij}}{t_{ij}} \right) - T_{ij} \right) \\ & = T \ln \left(\frac{T}{t} \right) - \sum_{ij} T_{ij} \ln \left(\frac{T_{ij}}{t_{ij}} \right). \end{aligned} \quad (32)$$

When Eq. (32) is augmented with the previously mentioned partial derivatives that minimize the link flow error we derive

$$\begin{aligned} \max. \quad & T \ln \left(\frac{T}{t} \right) - \sum_{ij} T_{ij} \ln \left(\frac{T_{ij}}{t_{ij}} \right) \\ & - \sum_{ij} \left(\lambda_{ij} 2 \left(\sum_a \left(V_a \cdot p_{ij}^a \right) - \left(\sum_a p_{ij}^a \left(\sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \right). \end{aligned} \quad (33)$$

This equation, when solved, yields the most likely O-D matrix of all of those matrices that

come equally close to matching the observed link flows.

It should be noted that the objective function of Eq. (33) is composed of two components. The first being the error between the field observed flows and the flows that satisfy flow continuity with minimum difference from observed flows. The second component represents the likelihood of an O-D matrix table. The objective is to find the O-D matrix with the maximum likelihood. In the case that the seed matrix is the optimum matrix the likelihood component resolves to zero.

Dynamic Formulations

The above formulations assume that the vehicles are assigned to all links simultaneously (i.e. a vehicle is present on all links along its path simultaneously). In order to address the dynamic nature of traffic, the analysis period can be divided into equally spaced time slices. Origin-destination demands are then indexed by the time slice they depart and the time slice they are observed on a link, as

$$\begin{aligned} \max. \quad & T_r \ln \left(\frac{T_r}{t_r} \right) - \sum_{rij} T_{rij} \ln \left(\frac{T_{rij}}{t_{rij}} \right) \\ & - \sum_{rij} \left(\lambda_{rij} 2 \left(\sum_{sa} \left(V_{sa} \cdot p_{rij}^{sa} \right) - \left(\sum_{sa} p_{rij}^{sa} \left(\sum_{rxy} T_{rxy} p_{rxy}^{sa} \right) \right) \right) \right). \end{aligned} \quad (34)$$

Where T_r is the total demand departing during time-slice r , t_r is the total seed matrix demand departing during time-slice r , T_{rij} is the traffic demand departing during time-slice r traveling between origin i and destination j , t_{rij} is the seed traffic demand departing during time-slice r traveling between origin i and destination j , λ_{rij} is the Lagrange multiplier for departure time-slice, origin, and destination combination rij , V_{sa} is the observed volume on link a during time slice s , and p_{rij}^{sa} is the probability of a demand between origin i and destination j during time-slice r is observed on link a during time-slice s . The solution of Eq. (34) is computationally extensive and has been demonstrated to not produce significantly better results than generating time-dependent O-D demands, as will be discussed.

Alternatively, the more common approach is to generate time-dependent O-D demands by solving Eq. (33) for each time-slice independently assuming that O-D demands can travel from the origin to destination zone within a time-slice (i.e. the trip travel time is less than the time-slice duration) without considering the interaction between time slices. This approach is computationally simpler and easier to implement and thus will be discussed in more detail. The formulation can be written as

$$\begin{aligned} \max. \quad & T_r \ln \left(\frac{T_r}{t_r} \right) - \sum_{ij} T_{rij} \ln \left(\frac{T_{rij}}{t_{rij}} \right) \\ & - \sum_{ij} (\lambda_{rij} \cdot 2 \left(\sum_a (V_a \cdot p_{rij}^a) \right. \\ & \left. - \left(\sum_a p_{rij}^a \left(\sum_{xy} T_{xy} p_{rxy}^a \right) \right) \right)) \quad \forall r. \end{aligned} \quad (35)$$

Here the T_{rij} are solved for independent of other time-slices. It should be noted, that the approach ignores the interaction of demands across various time-slices which is a valid assumption if the network is not over-saturated. However, if the network is oversaturated the assumption of time slice independence may not

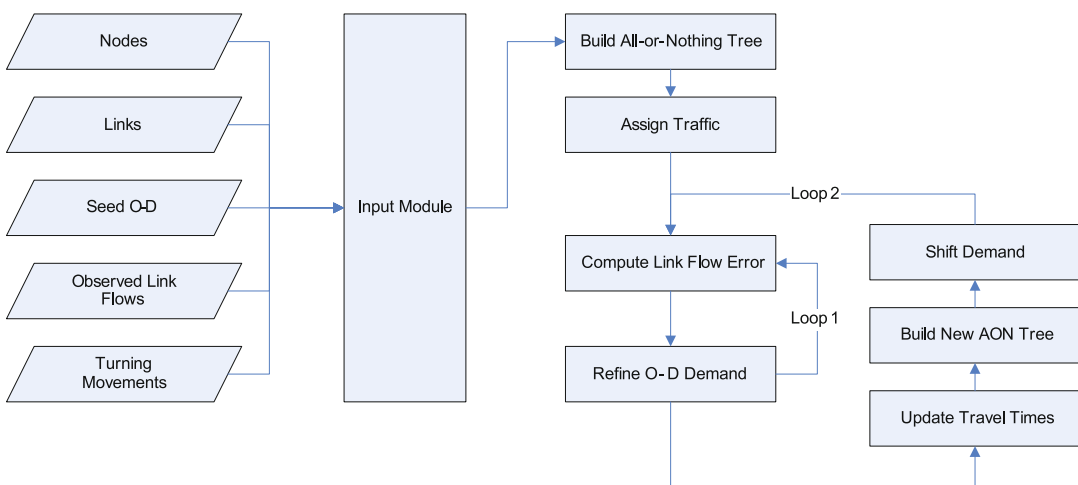
be valid. The duration of the time-slice should be selected such that steady-state conditions are achieved within a time-slice.

Solution Algorithms

The solution of the set of equations presented in (35) is hard given that the objective function is non-convex and that in many cases the p_{rij}^a are not available and thus the problem becomes to solve for T_{rij} and p_{rij}^a that maximize the objective function.

Here we present a numerical heuristic that solves the above formulation for large networks when the number of equations and unknowns becomes extremely computationally intensive. This special purpose equation solver has been developed and implemented in the QUEENSOD software. This solver fully optimizes the objective function of Eq. (35). The software has been shown to produce errors less than 1% for the range of values and derivatives being typically considered in the problem. A sample application of the QUEENSOD software is presented later in the paper, however, initially the heuristic approach is described.

The numerical solution begins by building a minimum path tree and performing an all-or-nothing traffic assignment of the seed matrix, as illustrated in Fig. 7. A relative or absolute link flow error is computed depending on user input.



Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks, Fig. 7 QueensOD heuristic O-D estimation approach (synthetic O-D estimator)

Using the link flow errors O-D adjustment factors are computed and utilized to modify the seed O-D matrix. The adjustment of the O-D matrix continues until one of two criteria are met, namely the change in O-D error reaches a user-specified minimum or the number of iterations criterion is met. If additional trees are to be considered, the model builds a new set of minimum path trees (loop 2) and shifts traffic gradually to the second minimum path tree. The minimum objective function for two trees is computed in a similar fashion as described for the single tree scenario. The process of building trees and finding the optimum solution continues until all possible trees have been explored. The proposed numerical solution ensures that in the case that the seed matrix is optimum no changes are made to the matrix. In addition, the use of the seed matrix as a starting point for the search algorithm ensures that the optimum solution resembles the seed matrix as closely as possible while minimizing the link flow error. In other words, the seed matrix biases the solution towards the seed matrix.

In order to demonstrate the applicability of the QUEENSOD software, a sample application to a 3500-link network of the Bellevue area in Seattle is presented. Other applications of the QUEENSOD software are described in detail in the literature (Dion et al. 2004a; Rakha et al. 1998). The O-D demand for the Bellevue network was calibrated to AM peak Single Occupancy Vehicle (SOV) and High Occupancy Vehicle (HOV) flows. The seed matrix was created using the standard four-step transportation planning process by applying the EMME/2 model. The Seattle network was converted from EMME/2 format to INTEGRATION format.

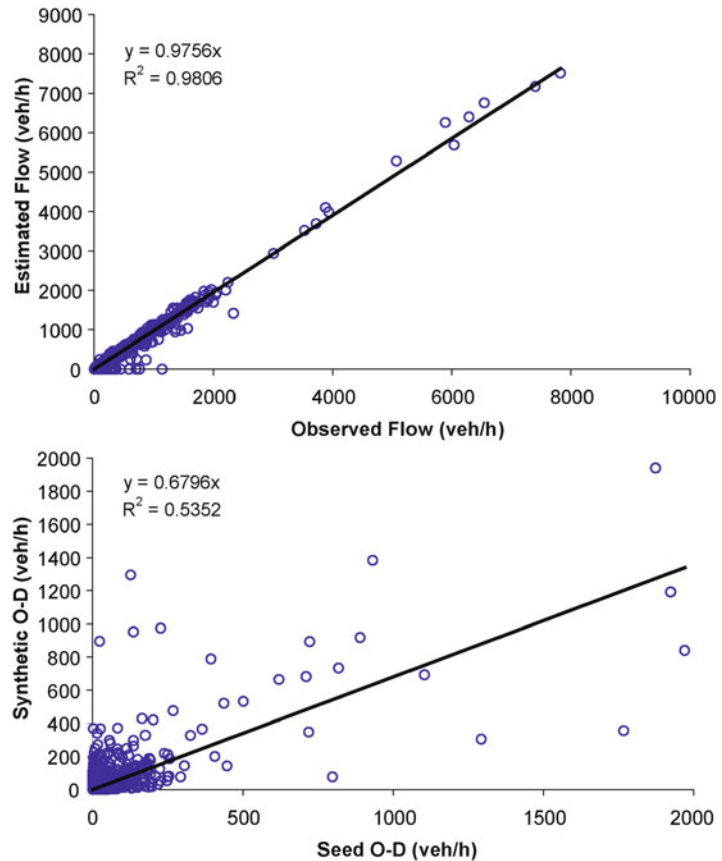
The calibration of the O-D demand to tube and turning movement counts was conducted using the QUEENSOD software using the planning trip distribution O-D matrix as the seed solution. The calibration resulted in a high level of consistency between estimated and field observed link flow counts (coefficient of determination of 0.98 between the estimated and observed flows), as illustrated in Fig. 8. Figure 8 demonstrates that in calibrating the O-D matrix to observed traffic counts, the trip distribution O-D matrix (seed matrix) was modified significantly (coefficient of

determination of 0.56 between trip distribution and synthetic O-D matrix). Consequently, it is evident that a modification of the trip distribution matrix was required in order to better match observed link and turning movement counts. It should be noted however, that the total number of trips was increased by only 4% as a result of the synthetic O-D calibration effort. Consequently, the illustrated calibration effort resulted in a significant modification of the trip table with minor modification to the total number of trips.

In addition to the above mentioned research, a significant number of problem formulations and applications have been documented in the literature. To name a few (in chronological order) Cascetta et al. (Cascetta and Marquis 1993) tested a method based on two generalized least squares estimators on the Brescia-Verona-Vicenza-Padua motorway in Italy. They found that the accuracy of their model depended heavily on the number of links with observed traffic counts. Van Aerde et al. (1993) introduced the QUEENSOD method and demonstrated its applicability on a 35-km section of Highway 401 in Toronto, Canada. Ashok (1996) evaluated the use of a Kalman filtering-based method, which was first presented by Okutani (1987) and estimates unobserved link traffic counts from observed link traffic counts. The method used was formulated by Ashok and Ben Akiva (1993) and Ashok (1996) and was evaluated using actual data from the Massachusetts Turnpike, Massachusetts, a stretch of I-880 near Hayward, California and a freeway encircling the city of Amsterdam, Netherlands. Later, Hellings and Van Aerde (1998) compared a least square error model and a least relative error model on a 35-km section of Highway 401 in Toronto, Canada. Zhou and Sachse (1997) compared the use of three different O-D estimators and on a motorway network in Europe. They concluded that the models, although characterized by different computational loads, produced satisfactory results. They also commented on the need to decide on locations of detectors and aggregation time intervals. Van Der Zijpp and Romph (1997) experimented their model on the Amsterdam Beltway. They tested their model using two different days worth of data and compared their model results with real and historical average data.

Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks,

Fig. 8 Example application of QUEENSOD to the Bellevue network in Seattle



While their model performed better in cases of accidents, the historical average data did, at least as good, in normal traffic. They stressed on the importance of correct modeling of the network and traffic flow characteristics for the production of good results. Kim et al. (2001) introduced a genetic algorithm based method to overcome the shortcoming of the bi-level programming method when there is a significant difference between target and true O-D matrices. They tested their model on a small network of 9 nodes. Bierlaire and Crittin (2004) formulated a least-square based method to overcome some of the shortcomings of the Kalman filter approach. They tested their method on a simple network as well as two real networks: a medium scale network, Central Artery Network, Boston, MA, and a large scale network, Irvine Network, Irvine, CA. Yun and Park developed a genetic algorithm based method with the purpose

of solving dynamic O-D matrices for large networks. They compared their model's results with the results of QUEENSOD, and they tested their method on the City of Hampton network using the PARAMICS microscopic traffic simulation software. Nie et al. (2005) developed a formulation that incorporates a decoupled path flow estimator in a generalized least squares framework with the objective of developing an efficient, simplified solution algorithm for realistic size networks. They tested their method on a small (9-node) and mid-size (100 nodes) network. Zhou and Mahmassani (2006) developed a multi-objective optimization framework for the estimation of the O-D matrices using automatic vehicle identification data. They tested their method on a simplified Irvine testbed network (31 nodes). Finally, Castillo et al. (2008) developed a method for the reconstruction and estimation of the trip matrix and path

flows based on plate scanning and link observations. They tested their method on the Nguyen–Dupuis Network, and concluded the superiority of plate scanning on link counts.

It should be noted at this point that the O-D estimation formulations and techniques that were presented and described in this section are heuristics and thus there is no mathematical proof that the algorithms converge to the unique optimum solution either in the static or dynamic context. While we have demonstrated that the solution matches the observed link flows for complex networks (Fig. 8, unfortunately the actual O-D demand is typically not available for real-life applications and thus it is not possible to measure how good the solution compares to the unique optimum O-D matrix.

Dynamic Estimation of Measures of Effectiveness

Dynamic assessment of traffic network performance requires the estimation of various measures of effectiveness in a dynamic context. This section provides a brief overview of the procedures for estimating delay, vehicle stops, and vehicle energy consumption and emissions.

Estimation of Delay

A key parameter in the dynamic assessment of traffic networks is the estimation of vehicle delay. The computation of delay requires the computation of travel times. Significant research has been conducted to develop analytical models for estimating delay especially at signalized intersections. Examples of such research efforts are provided for the interested reader (Akcelik and Roupail 1994; Brilon 1995; Brilon and Wu 1990; Cassidy and Han 1993; Cassidy et al. 1994; Catling 1977; Colyar and Roupail 2003; Cronje 1983a, b, 1986; Daganzo and Laval 2005; Daniel et al. 1996; Engelbrecht et al. 1996; Fambro and Roupail 1996; Fang et al. 2003; Flannery et al. 2005; Hagring et al. 2003; Krishnamurthy and Coifman 2004; Lawson et al. 1996; Li et al. 1994; Newell 1999; Roupail 1988; Roupail and Akcelik 1992; Tarko et al. 1993).

Roadway travel times can be computed for any given vehicle by providing that vehicle with a *time card* upon its entry to any roadway or link. Subsequently, this *time card* is retrieved when the vehicle leaves the roadway. The difference between these entry and exit times provides a direct measure of the roadway travel time experienced by each vehicle. The delay can then be computed as the difference between the actual and free-flow travel time.

Alternatively, vehicle delay can be computed microscopically every deci-second as the difference in travel time between travel at the vehicle's instantaneous speed and travel at free-flow speed, as

$$d(t_i) = \Delta t \left(1 - \frac{u(t_i)}{u_f} \right) \quad \forall i. \quad (36)$$

The summation of these instantaneous delay estimates over the entire trip provides an estimate of the total delay. This model has been validated against analytical time-dependent queuing models, shockwave analysis, the Canadian Capacity Guide, the Highway Capacity Manual (HCM), and the Australian Capacity Guide procedures (Dion et al. 2004a). The procedure has also been incorporated in the INTEGRATION traffic simulation software (Van Aerde and Rakha 2007) and utilized with second-by-second Global Positioning System (GPS) data (Dion et al. 2004b; Rakha et al. 2000).

Estimation of Vehicle Stops

Numerous researchers have dealt with the problem of estimating vehicle stops especially at signalized intersections. An important early contribution is attributed to Webster (1958), who generated stop and delay relationships by simulating uniform traffic flows on a single-lane approach to an isolated intersection. In particular, the equations that Webster (1958) derived have been fundamental to traffic signal setting procedures since their development. Later, Webster and Cobbe (1966) developed a formula for estimating vehicle stops at under-saturated intersections assuming random vehicle arrivals. Other models

were developed by Newell (1965) and Catling (1977). Catling adapted equations of classical queuing theory to oversaturated traffic conditions and developed a comprehensive queue length estimation procedure that captured the time-dependent nature of queues to be applied to both under-saturated and over-saturated conditions. In addition, Cronje (1983a, b, c, 1986) developed stop and delay equations by treating traffic flow through a fixed-time signal as a Markov process. The approach assumed that the number of queued vehicles at the beginning of a cycle could be expressed by a geometric distribution. These models, however, were not designed to account for the partial stops that vehicles may incur. Furthermore, the models that account for partial stops do not estimate vehicle partial and full stops for over-saturated conditions. A study by Rakha et al. (2001a) developed a procedure for estimating vehicle stops while accounting for partial stops, as

$$S(t_i) = \frac{u(t_i) - u(t_{i-1})}{u_f} \quad (37)$$

$$\forall i \ni u(t_i) < u(t_{i-1}).$$

The sum of these partial stops is also recorded. This sum, in turn, provides a very accurate explicit estimate of the total number of stops that are encountered along a roadway. Again the model can be implemented within a microscopic traffic simulation software or applied to second-by-second speed measurements using a GPS system.

Estimation of Vehicle Energy Consumption and Emissions

Estimating accurate mobile source emissions has gained interest among transportation professionals as a result of increasing environmental problems in large metropolitan urban areas. While current emission inventory models in the US, such as MOBILE and EMPAC, are capable of estimating large scale inventories, they are unable to estimate accurate vehicle emissions that result from operational-level projects. Alternatively, microscopic emission models are capable of assessing the impact of transportation projects on the environment and performing project level analyses.

Consequently, the focus of this discussion will be on these microscopic and also mesoscopic models. Two models that are emerging include the Comprehensive Modal Emissions Model (CMEM) and the Virginia Tech Microscopic (VT-Micro) model. These models are briefly described in terms of their structure, logic, and validity.

Comprehensive Modal Emission Model

The Comprehensive Modal Emissions Model (CMEM), which is one of the newest power demand-based emission models, was developed by researchers at the University of California, Riverside (Barth et al. 2000). The CMEM model estimates LDV and LDT emissions as a function of the vehicle's operating mode. The term "comprehensive" is utilized to reflect the ability of the model to predict emissions for a wide variety of LDVs and LDTs in various operating states (e.g., properly functioning, deteriorated, malfunctioning).

The development of the CMEM model involved extensive data collection for both engine-out and tailpipe emissions of over 300 vehicles, including more than 30 high emitters. These data were measured at a second-by-second level of resolution on three driving cycles, namely: the Federal Test Procedure (FTP), US06, and the Modal Emission Cycle (MEC). The MEC cycle was developed by the UC Riverside researchers in order to determine the load at which a specific vehicle enters into fuel enrichment mode. CMEM predicts second-by-second tailpipe emissions and fuel consumption rates for a wide range of vehicle/technology categories. The model is based on a simple parametrized physical approach that decomposes the entire emission process into components corresponding to the physical phenomena associated with vehicle operation and emission production. The model consists of six modules that predict engine power, engine speed, air-to-fuel ratio, fuel use, engine-out emissions, and catalyst pass fraction. Vehicle and operation variables (such as speed, acceleration, and road grade) and model calibrated parameters (such as cold start coefficients, engine friction factor) are utilized as input data to the model.

Vehicles were categorized in the CMEM model based on a vehicle's total emission contribution. Twenty-eight vehicle categories were

constructed based on a number of vehicle variables. These vehicle variables included the vehicle's fuel and emission control technology (e.g. catalyst and fuel injection), accumulated mileage, power-to-weight ratio, emission certification level (tier0 and tier1), and emitter level category (high and normal emitter). In total 24 normal vehicle and 4 high emitter categories were considered (Barth et al. 2000).

The Virginia Tech Microscopic Energy and Emission Model (VT-Micro Model) The VT-Micro emission models were developed from experimentation with numerous polynomial combinations of speed and acceleration levels. Specifically, linear, quadratic, cubic, and fourth degree combinations of speed and acceleration levels were tested using chassis dynamometer data collected at the Oak Ridge National Laboratory (ORNL). The final regression model included a combination of linear, quadratic, and cubic speed and acceleration terms because it provided the least number of terms with a relatively good fit to the original data (R^2 in excess of 0.92 for all measures of effectiveness [MOE]). The ORNL data consisted of nine normal-emitting vehicles including six light-duty automobiles and three light-duty trucks. These vehicles were selected in order to produce an average vehicle that was consistent with average vehicle sales in terms of engine displacement, vehicle curb weight, and vehicle type. The data collected at ORNL contained between 1,300 to 1,600 individual measurements for each vehicle and MOE combination depending on the vehicle's envelope of operation (Ahn et al. 2002).

This method has a significant advantage over emission data collected from a few driving cycles because it is difficult to cover the entire vehicle operational regime with only a few driving cycles. Typically, vehicle acceleration values ranged from -1.5 to 3.7 m/s^2 at increments of 0.3 m/s^2 (-5 to 12 ft/s^2 at 1-ft/s^2 increments). Vehicle speeds varied from 0 to 33.5 m/s (0 to 121 km/h or 0 to 110 ft/s) at increments of 0.3 m/s (Ahn et al. 2002).

The model had the problem of overestimating HC and CO emissions especially for high acceleration levels. Since this problem arose from the fact that the sensitivity of the dependent variables to the positive acceleration levels is significantly

different from that for the negative acceleration levels, a two-regime model for positive and negative acceleration regimes was developed as (Ahn et al. 2002; Rakha et al. 2004a).

$$\ln(MOE_e) = \begin{cases} \sum_{i=0}^3 \sum_{j=0}^3 (L_{i,j}^e \times u^i \times a^j) & \text{for } a \geq 0 \\ \sum_{i=0}^3 \sum_{j=0}^3 (M_{i,j}^e \times u^i \times a^j) & \text{for } a < 0. \end{cases} \quad (38)$$

Where MOE_e is the instantaneous fuel consumption or emission rate (ml/s or mg/s); $K_{i,j}^e$ is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j"; $L_{i,j}^e$ is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j" for positive accelerations; $M_{i,j}^e$ is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j" for negative accelerations; u is the instantaneous speed (km/h); and a is the instantaneous acceleration rate (km/h/s).

Additionally, the VT-Micro model was expanded by including data from 60 light-duty vehicles (LDVs) and trucks (LDTs). Statistical clustering techniques were applied to group vehicles into homogeneous categories using classification and regression tree (CART) algorithms. The 60 vehicles were classified into five LDV and two LDT categories (Rakha and Ahn 2004). In addition, HE vehicle emission models were constructed using second-by-second emission data. In constructing the models, HEVs are classified into four categories for modeling purposes. The employed HEV categorization was based on the comprehensive modal emission model (CMEM) categorization. The first type of HEVs has a chronically lean fuel-to-air ratio at moderate power or transient operation, which results in high emissions in NO. The second type has a chronically rich fuel-to-air ratio at moderate power, which results in high emissions in CO. The third type is high in HC and CO. The fourth type has a chronically or transiently poor catalyst performance, which results in high emissions in HC, CO, and NO. Each model for each category was constructed within the VT-Micro modeling framework. The HE vehicle model was found to estimate vehicle

emissions with a margin of error of 10% when compared to in-laboratory bag measurements (Ahn et al. 2004). Furthermore, all the models were incorporated into the INTEGRATION software, and made it possible to evaluate the environmental impacts of operational level transportation projects (Park and Rakha 2006).

Use of Technology to Enhance System Performance

Due to the recent extensive developments within the fields of artificial intelligence, communications, and computation algorithms, transportation and traffic engineers' goals have evolved. As mentioned earlier in the paper, current spatio-temporal distribution of trips is far from being optimum, either with respect to driver satisfaction and/or network performance. A part of the contemporary DTA research is directed towards influencing, as opposed to modeling, dynamic spatio-temporal trip distributions. Advanced Traveler Information Systems (ATISs) are definitely the main tool for such influence, and understanding driver behavior is critical to the design and implementation of such systems. Research which is directly related to the possibility of enhancing system performance through the use of technology may be categorized in the following main areas of research:

- Validation of models, lab experiments and real world behavior, which is the area concerned with verifying the different theories and their implicit assumptions with regards to real-life situations. Due to the extreme complexity and questionable possibility of this task, several attempts have been made to verify the models with respect to lab experiments rather than the real world behavior. Moreover, comparison and verification of spatial and temporal transferability of the models might as well fall within this area. Examples of current literature include the works of Chang and Mahmassani (1988) and Mahmassani and Jou (2000).
- Calibration of algorithms and models, which as the name suggests, is the area related to the calibration of the algorithms and model parameters. This also entails spatial and temporal calibration, for certain models and/or parameters might only be valid for certain locations and time periods rather than others. Examples of current literature include the works of Chang and Mahmassani (1988) and Rakha and Arafah (2007).
- Real time deployment, which focuses on the possibility of deploying DTA models into the real world. This area of research is concerned with developing deployable DTA algorithms. Current literature states that although "a mathematically tractable analytical model that is adequately sensitive to traffic realism vis-à-vis real-time operation is still elusive," yet even with currently available models there is a tradeoff between solution accuracy and computational efficiency. Other real-time deployment issues include computational tractability; consistency checking; model robustness, stability, and error and fault tolerance; and demand estimation and prediction (Peeta and Ziliaskopoulos 2001). Examples of current literature include the works of Mahmassani et al. (1998a, b, c; Mahmassani and Peeta 1993), Ben-Akiva et al. (1997, 1998b), Mahmassani and Peeta (1992, 1993, 1995), Peeta and Mahmassani (1995a), Hawas (1995), Hawas et al. (Hawas and Mahmassani 1997), Hawas and Mahmassani (1995; Hawas et al. 1997), Cantarell and Cascetta (1995), Anastassopoulos (2000), and Jha et al. (1998).
- Issues of uncertainty, which is, as mentioned earlier, a fundamental feature in most transportation phenomena. Uncertainty can be represented in trip makers' knowledge of different route travel times, in the compliance rates of drivers to information, in the accuracy of the disseminated control information, in the driver's perception of disseminated information reliability, in the controller's predicted and/or refined dynamic travel times and/or O-D matrices, among others. Uncertainty-related research issues have been addressed through several approaches, like stochastic modeling, fuzzy control, and reliability indices. Examples of current literature include the

works of: Birge and Ho (1993), Peeta and Zhou (Kerner and Klenov 2006; Peeta and Bulusu 1999), Cantarell and Cascetta (1995), Ziliaskopoulos and Waller (Ziliaskopoulos 2000), Waller and Ziliaskopoulos (2006), Waller (2000), Peeta and Yu (2006), Jha et al. (1998), Peeta and Paz (2006).

- DTA control, which is the area of research concerned with modifying how trips are distributed on the network. Research within this area focuses on capturing current network performance, and works on modifying the system elements, such as drivers route, and/or departure time selection, as well as mode choice (possibly through pricing and information dissemination); and traffic management (primarily through signal operation), in order to optimize system performance. Examples of current literature include the works of Peeta and Paz (2006).
- Realism of other system characteristics, which is the research area concerned with capturing other system realities that are not considered in current available literature. Examples of such realities may include (Peeta and Ziliaskopoulos 2001),
 - Person rather than driver assignment. It is an undeniable fact that many people tend to make their mode choices based on daily, real-time decisions, i.e. this is a dynamic and not a static process. It is further anticipated that with the current (and predicted) maturity of information technology within the transportation arena, would require explicit modeling within DTA models.
 - The effect of interaction between the different vehicle classes and road infrastructure. It is beyond doubt that certain vehicle classes (such as trucks and busses for example) will not be able to comply with certain diversion-requesting disseminated information, due to road infrastructure constraints. However, in other occasions, these vehicle classes might be able to divert routes, yet with travel time penalties (example if the turning radius was inadequate) that might not only affect these vehicle classes, but all other diverting vehicles as well.
 - Capturing latest traffic control technology and strategies. Traffic control technology

and strategies have been rapidly developing during the past couple of decades. Examples of this include transit signal preemption, real-time adaptive signal traffic control, electronic toll collection, etc. For efficient DTA control, DTA algorithms should be able to sufficiently capture and consider them.

Examples of current literature include the works of Ran and Boyce (1996), Peeta et al. (Peeta and Yang 2000), Ziliaskopoulos and Waller (Ziliaskopoulos 2000), Dion and Rakha (Dion et al. 2004b), Sivananden et al. (2003), Rakha et al. (2000, 2005b), Rakha and Zhang (2004a).

Road Pricing

A number of approaches have been and continue to be considered in an attempt to reduce urban congestion. These approaches include supply and demand management strategies. Supply management strategies are addressed through the building of new infrastructure, expanding the existing infrastructure, or by increasing the utilization of existing infrastructure through the introduction of traffic management systems. Alternatively, demand strategies focus on either distributing the existing demand more evenly in space over the network or distributing the existing demand more evenly by spreading the peak period over a longer time. One of the emerging demand management strategies for enhancing traffic system performance is the use of electronic road pricing.

Road pricing entails the charging of the user of a road for the use of the facility. The road charges include fuel taxes, license fees, parking taxes, tolls, and congestion charges, including those which may vary by time-of-day, by the specific vehicle type being used. Road pricing has two distinct objectives: revenue generation, usually for road infrastructure financing, and congestion pricing for demand management purposes. Toll roads are the typical example of revenue generation. Changes for using high occupancy toll (HOT) lanes or urban tolls for entering a restricted area of a city are typical examples of using road pricing for congestion management purposes.

In modeling route choice within the context of road pricing one has to consider at least two main types of route choices, one which is tolled and one which is not. An example of such a situation is a multi-lane freeway that splits at several locations to provide 3 lanes that are free of toll and 1 lane that involves a toll. While these are some unique traffic engineering issues associated with the weaving that can take place, when there are multiple breaks in the structure that divides the toll lanes from the free ones, many important equilibrium issues can already be analyzed by looking at single diverge sections. If traffic demand is low, and the distance and speed limits are similar on both types of lanes, there exists little incentive for drivers to utilize the toll lane alternative. However, as the traffic demand on the toll alternative increases, its speed will decrease. This will make its travel time become longer, making the free lanes less attractive. This sets up a condition in which it may become attractive to some, but not necessarily all drivers, to start utilizing the toll lane. The timing of when drivers start to use the toll lanes and the number of drivers who elect to use the toll lane depends primarily on two factors, namely the “value of the toll” and the “value of time” of the drivers involved.

The “value of the toll” and “value of time” are closely related quantities, as a higher toll value accompanied with a higher value of time will result in the same perceived disutility and therefore path impedance. Consequently, it is ratio of the “value of the toll” to the “value of time” that is most important. In addition, one should also note that for the same toll value, those with the highest “value of time” will be least influenced by the toll. They will therefore also be the first ones to start using the toll lanes when the free lanes become busy and slower. In contrast, those with the lowest value of time perceive the toll as having disutility or travel time equivalency. They will therefore be the last ones to switch to the toll lane, if at all.

The decisions, of those drivers who have a lower value of time, are also influenced by the decisions of the drivers with a higher value of time. Specifically, any switch over of the latter drivers from the free lanes to the toll lanes will leave the free lanes less congested for those with a lower value of time who stay in the free lanes.

This may leave it undesirable for those with a low value of time to also switch to the toll lane. However, if the percent of drivers with a high value of time is relatively small, or if the total flow A to B is very high, even some of those with a low value of time will perceive themselves as being better off if they use the toll lanes. The issue of toll lane usage therefore requires one to answer two interrelated questions, namely: (a) what group or sub-population of drivers will utilize the toll road, and (b) what fraction of this group or sub-population will utilize the toll road. Consequently, extensive research is currently underway to model driver route selection under various tolling conditions. Examples of some research efforts on this topic include (Ben-Akiva et al. 1993; Mekky 1995, 1996, 1998; Munnich et al. 2007). This list is by no means comprehensive but does provide some applications on this emerging topic.

Related Transportation Areas

Research within the following two transportation areas definitely precedes DTA research. However, their significance to the DTA field is based on the fact that DTA theories are mostly dependent on older theories stemming from these two areas. Hence, advances within these two areas could probably significantly affect the advances within the DTA arena.

- Traffic flow models encompass the mathematical representation, or perhaps simulation of the traffic flow characteristics, such as modeling traffic flow propagation, queue spillbacks, lane-changing, signal operation, travel time computation, etc. are crucial in determining driver expectations and behavior. In addition, these are also fundamental in the calculations of travel times, which are vital in the combined problem of departure time and route choice. The quantity of research available in this area is probably as big as the quantity of research done in the area of DTA all together, if not even more. However, as mentioned earlier, all of the research done within this area has direct influence on the realization of the traffic flow models, which are also used within the DTA models.

- Planning applications, which in spite of being a quite under-researched area at the moment, is a vitally important one. There is no doubt DTA models are superior to static models, hence, it is probably only a matter of time before the industry abandons static models for dynamic models. “Dynamic models are simply the natural evolution in the transportation field that like any other new effort suffers from early development shortcomings” (Peeta and Ziliaskopoulos 2001). Examples of current literature within this area of research include the works of Li (2001), Friesz et al., Waller (Waller 2000), Waller and Ziliaskopoulos (2006), Ziliaskopoulos and Waller (2000), Ziliaskopoulos and Wardell (2000).

Future Directions

Future research challenges and directions include:

- Enhance traffic flow modeling and driver behavior modeling. These include the modeling of person as opposed vehicle route choices, the separation of driver and vehicle within the traffic modeling framework, the explicit modeling of vehicle dynamics, enhancing car following, lane-changing, and routing behavior.
- Develop more efficient algorithms that would be suitable for real-time deployment, without making any compromises in the computational accuracy, i.e. without trading-off the solution accuracy for the computational efficiency. In precise, without compromising any dimension of the traffic flow theory, nor driver behavior assumptions. As a matter of fact, further research should be done to capture more of the traffic, as well as the driver behavior theory. Hence, this should help in improving the realism of the available DTA models.
- Conduct more research on the driver behavior theory. Especially, since human factors cognitive research has significantly improved in the previous couple of decades, then modeling driver behavior from this perspective might lead to valuable outcomes.
- Critical examination of the validity of network equilibrium as a framework for network flow analysis (Nakayama et al. 2001). Many of the current algorithms are based on the assumption that drivers become rational and homogeneous with learning. Hence, resulting in network equilibrium. A number of recent research efforts suggest that some drivers remain less rational, and heterogeneous drivers make up the system; drivers’ attitudes toward uncertainty become bipolar; and some drivers are sometimes deluded. Further research is required to characterize and model such behavior.
- Validate current models by comparing current model outputs with real world experiments, and possibly with controlled lab experiments (as mid-way experiments before conducting real world evaluations).
- Enhance traffic modeling tools within DTA models to capture the effect of diversion compliance of different vehicle modes (especially heavy vehicles) to more geometrically restrictive highways.
- Possibly calibrating hybrid fuzzy-stochastic models and comparing results to traditional models. According to the work done by Chen (2000), probabilistic methods are better than possibility-based methods if sufficient information is available, on the other hand, possibility-based methods can be better if little information is available. However, when there is little information available about uncertainties, a hybrid method may be optimum.
- Conduct further research on the dynamic synthetic O-D estimation from link flow measurements and vehicle probe data. Further research is required to quantify the impact of erroneous or missing data on the accuracy of O-D estimates.
- Conduct further research on the temporal distribution of demand, analyzing and modeling it. Then, including the estimation and forecast of time-dependent demand within the planning process, in addition to the dynamic traffic management and control processes. This should, hopefully, help to fill-in the gap between the three mentioned processes.
- Incorporating person assignment, rather than mode assignment in the DTA and planning models, for as mentioned earlier, mode split is currently more of a daily real-time dynamic, rather than a static decision.

- Research is needed to develop models for driver behavior to different ATIS systems: (types and/or scenarios). Current literature is mainly based on stated preference surveys, which are known for their lack of accuracy. Before the deployment of ATIS systems, stated preference surveys were the best approach for prediction and modeling drivers' reactions. However now, after the deployment of many ATIS systems, more research is needed to capture the actual (possibly revealed) drivers' behavior, rather than the stated behavior.
- Develop approaches that are capable of realistically capturing traffic flow, traffic control, and their interactions; and simultaneously optimizing traffic flow routing and control. In other words, developing algorithms that are actually capable of capturing real-time driver behavior, and are able to control it, in order to improve network performance.

Appendix

Term Abbreviations

ANN	Artificial Neural Networks
ATIS	Advanced Traveler Information System
AVI	Automatic Vehicle Identification
AVL	Automatic Vehicle Location
DTA	Dynamic Traffic Assignment
FHWA	Federal Highway Administration
GA	Genetic Algorithm
GPS	Global Positioning System
HCM	Highway Capacity Manual
HOV	High Occupancy Vehicle
ITS	Intelligent Transportation Systems
LDV	Light Duty Vehicle
LMC	Link Marginal Cost
LP	Linear Programming
MOE	Measure of Effectiveness
NLP	Non-Linear Programming
O-D	Origin – Destination
PMC	Path Marginal Cost
SO	System Optimum
SOV	Single Occupancy Vehicle
TT	Travel Time
UE	User Equilibrium
VMS	Variable Message Sign

Variable Definitions

v_i	Traffic volume on route i
N	Set of network nodes
A	Set of network arcs (links)
R	Set of origin centroids
S	Set of destination centroids
k_{rs}	Set of paths connecting O-D pair ($r-s$); $r \in R, s \in S$
x_a	Flow on arc (a)
x_b	Flow on arc (b)
t_a	Travel time on arc (a)
t_b	Travel time on arc (b)
f_k^{rs}	Flow on path (k) connecting O-D pair ($r-s$)
f_l^{mn}	Flow on path (l) connecting O-D pair ($m-n$)
c_k^{rs}	Travel time on path (k) connecting O-D pair ($r-s$)
q_{rs}	Trip rate between origin (r) and destination (s)
$\delta_{a,k}^{rs}$	Indicator variable, = 1 if arc (a) is on path (k) between O-D pair ($r-s$), and 0 otherwise
x	Vector of flows on all arcs, $D(\dots, x_a, \dots)$
t	Vector of travel times on all arcs, $D(\dots, t_a, \dots)$
f^{rs}	Vector of flows on all paths connecting O-D pair $r-s$, $= (\dots, f_k^{rs}, \dots)$
f	Matrix of flows on all paths connecting all O-D pairs, $= (\dots, f^{rs}, \dots)$
c^{rs}	Vector of travel times on all paths connecting O-D pair $r-s$, $= (\dots, c^{rs}, \dots)$
c	Matrix of travel times on all paths connecting all O-D pairs, $D(\dots, c^{rs}, \dots)$
q	Origin-destination matrix (with elements $= q_{rs}$)
Δ^{rs}	Link-path incidence matrix (with $\delta_{a,k}^{rs}$ elements) for O-D pair $r-s$, as discussed below
Δ	Matrix of link-path incidence matrices (for all O-D pairs), $= (\dots, \Delta^{rs}, \dots)$
z	Objective function
L	Lagrange (transformation of the) objective function
u_{rs}	Dual variable associated with the flow conservation constraint for O-D pair ($r-s$)

$t_{i,k}$	Observed average travel time along link i within the k th sampling interval	l_{rij}	Lagrange multiplier for departure time-slice, origin, and destination combination (rij)
$\tilde{t}_{i,k}$	Smoothed average travel time along link i in the k th sampling interval	V_{sa}	Observed volume on link (a) during time-slice (s)
$s_{i,k}^2$	Variance of the observed travel times relative to the observed average travel time in the k th sampling interval	p_{rij}^{sa}	Probability of (a) demand between origin (i) and destination (j) during time-slice (r) is observed on link (a) during time-slice (s)
$\tilde{s}_{i,k}^2$	Variance of the observed travel times relative to the smoothed travel time in the k th sampling interval	$d(t_i)$	Vehicle delay at time (t_i)
$n_{i,k}$	Number of valid travel time readings on link i in the k th sampling interval	$u(t_i)$	Vehicle instantaneous speed at time (t_i)
α	Exponential smoothing factor that varies as a function of the number of observations $n_{i,k}$ within the sampling interval	u_f	Free-flow speed
β	Constant that varies between 0 and 1	$S(t_i)$	Vehicle full and partial stops at time (t_i)
T_{ij}	Number of trips between production zone i and attraction zone j	MOE_e	Instantaneous fuel consumption or emission rate
P_i	Number of trip productions from the origin zone	$K_{i,j}^e$	Model regression coefficient for MOE (e) at speed power (i) and acceleration power (j)
A_j	Number of trip attractions to the destination zone	$L_{i,j}^e$	Model regression coefficient for MOE (e) at speed power (i) and acceleration power (j) for positive accelerations
F_{ij}	Impedance factor between production zone i and attraction zone j	$M_{i,j}^e$	Model regression coefficient for MOE (e) at speed power (i) and acceleration power (j) for negative accelerations
K_{ij}	Socio-economic adjustment factor for trips between production zone i and attraction zone j	u	Vehicle instantaneous speed
c_{ij}	Generalized cost of inter-zonal travel between production zone i and attraction zone j	a	Vehicle instantaneous acceleration rate
t_{ij}	Prior information on the number of trips between production zone i and attraction zone j	q	Traffic stream flow (veh/h)
V_a	Traffic flow on link (a)	k	Traffic stream density (veh/km)
V'_a	Complementary traffic flow on link (a)	u	Traffic stream space-mean speed (km/h)
p_{ij}^a	Probability of traffic flow between origin (i) and destination (j) to use link (a)	u_f	Expected traffic stream free-flow speed (km/h)
T_r	Total demand departing during time-slice (r)	u_c	Expected traffic stream speed-at-capacity (km/h)
t_r	Total seed matrix demand departing during time-slice (r)	k_j	Expected traffic stream jam density (veh/km)
T_{rij}	Traffic demand departing during time-slice (r) traveling between origin (i) and destination (j)	q_c	Expected traffic stream capacity (veh/km)
t_{rij}	Seed traffic demand departing during time-slice (r) traveling between origin (i) and destination (j)	c_1	Model coefficient (km/veh)
		c_3	Model constant (h/km ² -veh)
		c_2	Model constant (h ⁻¹)

Bibliography

- Abdel-Aty MA, Kitamura R, Jovanis PP (1997) Using stated preference data for studying the effect of advanced traffic information on drivers' route choice. Transp Res Part C Emerg Technol 5(1):39–50

- Abdelfatah AS, Mahmassani HS (2001) A simulation-based signal optimization algorithm within a dynamic traffic assignment framework. In: IEEE intelligent transportation systems proceedings, IEEE conference on intelligent transportation systems, Proceedings, ITSC 2001, Oakland
- Abdelghany KF, Mahmassani HS (2001) Dynamic trip assignment-simulation model for intermodal transportation networks. *Transp Res Rec* 1771:52–60
- Abdelghany KF, Valdes DM et al (1999) Real-time dynamic traffic assignment and path-based signal coordination: application to network traffic management. *Transp Res Rec* 1667:67–76
- Abdelghany AF, Abdelghany KF et al (2000) Dynamic traffic assignment in design and evaluation of high-occupancy toll lanes. *Transp Res Rec* 1733:39–48
- Abdulhai B, Porwal H, Recker W (2002) Short-term freeway traffic flow prediction using genetically optimized time delay-based neural networks. *ITS J Intell Transp Syst J* 7(1):3–41
- Ahmed M, Cook AR (1982) Analysis of freeway traffic time series data by using Box-Jenkins techniques. *Transp Res Rec* 722:1–9
- Ahn K, Rakha H et al (2002) Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *J Transp Eng* 128(2):182–190
- Ahn K, Rakha H et al (2004) Microframework for modeling of high-emitting vehicles. *Transp Res Rec* 1880:39–49
- Akcelik R, Roupail NM (1994) Overflow queues and delays with random and platooned arrivals at signalized intersections. *J Adv Transp* 28(3):227–251
- Allen RW, Stein AC, Rosenthal TJ, Ziedman D, Torres JF, Halati A (1991) Human factors simulation investigation of driver route diversion and alternate route selection using in-vehicle navigation systems. In: Vehicle navigation & information systems conference, Dearborn, 20–23 Oct 1991. Proceedings Part 1 (of 2) Society of Automotive Engineers. SAE, Warrendale, pp 9–26
- Anastassopoulos I (2000) Fault-tolerance and incident detection using Fourier transforms. Purdue University, Westlafayette
- Arafeh M, Rakha H (2005) Genetic algorithm approach for locating automatic vehicle identification readers. In: IEEE intelligent transportation system conference, Vienna, 2005. Proceedings ITSV'05 IEEE intelligent conference on transportations systems, pp 1153–1158
- Arnott R, de Palma A, Lindsey R (1991) Does providing information to drivers reduce traffic congestion? *Transp Res Part A (General)* 25A(5):309
- Arrow KJ (1951) Alternative approaches to the theory of choice in risk-taking situations. *Econometrica* 19(4):404–437
- Ashok K (1996) Estimation and prediction of time-dependent origin-destination flows. PhD thesis, Massachusetts Institute of Technology, Boston
- Ashok K, Ben-Akiva ME (1993) Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. In: Daganzo CF (ed) 12th international symposium on transportation and traffic theory. Elsevier, New York, pp 465–484
- Ashok K, Ben-Akiva ME (2000) Alternative approaches for realtime estimation and prediction of time-dependent origin destination flows. *Transp Sci* 34(1):21–36
- Balakrishna R, Koutsopoulos HN et al (2005) Simulation-based evaluation of advanced traveler information systems. *Transp Res Rec* 1910:90–98
- Barth M, An F et al (2000) Comprehensive modal emission model (CMEM): version 2.0 user's guide. University of California, Riverside
- Bell M, Iida Y (1997) Transportation network analysis. Iida Y translator. Wiley, Chichester/New York
- Ben-Akiva M, Kroes E et al (1992) Real-time prediction of traffic congestion. Vehicle Navigation and Information Systems, IEEE, New York
- Ben-Akiva M, Bolduc D et al (1993) Estimation of travel choice models with randomly distributed values of time. *Transp Res Rec* 1413:88–97
- Ben-Akiva M, Bierlaire M, Bottom J, Koutsopoulos H et al (1997) Development of a route guidance generation system for real-time application. In: 8th international federation of automatic control symposium on transportation systems, Chania, 16–18 June 1997
- Ben-Akiva M, Bierlaire M et al (1998a) DynaMIT: a simulation based system for traffic prediction. DACCORD Short Term Forecasting Workshop, Delft, February 1998
- Ben-Akiva MMB, Koutsopoulos H, Mishalani R (1998b) DynaMIT: a simulation-based system for traffic prediction. DACCORD Short Term Forecasting Workshop, Delft
- Bierlaire M, Crittin F (2004) An efficient algorithm for real-time estimation and prediction of dynamic OD tables. *Oper Res* 52(1):116–127
- Birge JR, Ho JK (1993) Optimal flows in stochastic dynamic networks with congestion. *Oper Res* 41(1):203–216
- Bolland JD, Hall MD et al (1979) SATURN: simulation and assignment of traffic in urban road networks. In: International conference on traffic control systems, Berkeley
- Boyce DE, Ran B, Leblanc LJ (1995) Solving an instantaneous dynamic user-optimal route choice model. *Transp Sci* 29(2):128–142
- Braess D (1968) Über ein Paradoxon der Verkehrsplanung. *Unternehmensforschung* 12:258–268
- Brilon W (1995) Delays at oversaturated unsignalized intersections based on reserve capacities. *Transp Res Rec* 1484:1–8
- Brilon W, Wu N (1990) Delays at fixed-time traffic signals under time-dependent traffic conditions. *Traffic Eng Control* 31(12):8
- Burrell JE (1968) Multipath route assignment and its application to capacity-restraint. In: Fourth international symposium on the theory of traffic flow, Karlsruhe
- Burrell JE (1976) Multipath route assignment: a comparison of two methods. In: Florian M (ed) Traffic equilibrium methods. Lecture notes in economics and mathematical systems, vol 118. Springer, New York, pp 210–239

- Busemeyer JR, Townsend JT (1993) Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychol Rev* 100(3):432
- Byung-Wook Wie, TRL, Friesz TL, Bernstein D (1995) A discrete time, nested cost operator approach to the dynamic network user equilibrium problem. *Transp Sci* 29(1):79–92
- Cantarella GE, Cascetta ES (1995) Dynamic processes and equilibrium in transportation networks: towards a unifying theory. *Transp Sci* 29(4):305–329
- Carey M (1986) Constraint qualification for a dynamic traffic assignment model. *Transp Sci* 20(1):55–58
- Carey M (1987) Optimal time-varying flows on congested networks. *Oper Res* 35(1):58–69
- Carey M (1992) Nonconvexity of the dynamic traffic assignment problem. *Transp Res Methodol* 26B(2):127
- Carey M, Subrahmanian E (2000) An approach to modeling time-varying flows on congested networks. *Transp Res Methodol* 34B(3):157
- Cascetta E, Marquis G (1993) Dynamic estimators of origin-destination matrices using traffic counts. *Transp Sci* 27(4):363–373
- Cassidy MJ, Han LD (1993) Proposed model for predicting motorist delays at two-lane highway work zones. *J Transp Eng* 119(1):27–42
- Cassidy MJ, Rudjanakanoknad J (2005) Increasing the capacity of an isolated merge by metering its on-ramp. *Transp Res Part B Methodol* 39(10):896–913
- Cassidy MJ, Windover JR (1995) Methodology for assessing dynamics of freeway traffic flow. *Transp Res Rec* 1484:73–79
- Cassidy MJ, Son Y et al (1994) Estimating motorist delay at two-lane highway work zones. *Transp Res Part A Policy Pract* 28(5):433–444
- Castillo E, Menendez JM, Jimenez P (2008) Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. *Transp Res Part B Methodol* 42(5):455–481
- Catling I (1977) A time-dependent approach to junction delays. *Traffic Eng Control* 18(11):520–523, 526
- Chang GL, Mahmassani HS (1988) Travel time prediction and departure time adjustment behavior dynamics in a congested traffic system. *Transp Res Part B Methodol* 22B(3):217–232
- Chang GL, Tao X (1999) Integrated model for estimating time varying network origin-destination distributions. *Transp Res Part A Policy Pract* 33(5):381–399
- Chen SQ (2000) Comparing probabilistic and fuzzy set approaches for design in the presence of uncertainty. In: *Aerospace and ocean engineering*. PhD, Polytechnic Institute and State University, Blacksburg
- Chiu YC, Mahmassani HS (2001) Toward hybrid dynamic traffic assignment-models and solution procedures. In: *IEEE intelligent transportation systems proceedings, IEEE conference on intelligent transportation systems, Proceedings, ITSC 2001, Oakland*
- Coifman B (1998) New algorithm for vehicle reidentification and travel time measurement on freeways. In: *Proceedings of the 1998 5th international conference on applications of advanced technologies in transportation engineering*. ASCE, Reston
- Coifman B, Banerjee B (2002) Vehicle reidentification and travel time measurement on freeways using single loop detectors-from free flow through the onset of congestion. In: *Proceedings of the seventh international conference on: applications of advanced technology in transportation*, Cambridge, 5–7 Aug 2002. *Proceedings of the international conference on applications of advanced technologies in transportation engineering*. American Civil Engineers
- Coifman B, Cassidy M (2001) Vehicle reidentification and travel time measurement, Part I: Congested freeways. In: *IEEE intelligent transportation systems proceedings, Conference IEEE on intelligent transportation systems, Proceedings, ITSC 2001, Oakland*
- Coifman B, Ergueta E (2003) Improved vehicle reidentification and travel time measurement on congested freeways. *J Transp Eng* 129(5):475–483
- Colyar JD, Roupail NM (2003) Measured distributions of control delay on signalized arterials. *Transp Res Rec* 1852:1–9
- Cremer M, Keller H (1987) New class of dynamic methods for the identification of origin-destination flows. *Transp Res Part B Methodol* 21(2):117–132
- Cronje WB (1983a) Analysis of existing formulas for delay, overflow, and stops. *Transp Res Rec* 905:89–93
- Cronje WB (1983b) Derivation of equations for queue length, stops, and delay for fixed-time traffic signals. *Transp Res Rec* 905:93–95
- Cronje WB (1983c) Optimization model for isolated signalized traffic intersections. *Transp Res Rec* 905:80–83
- Cronje WB (1986) Comparative analysis of models for estimating delay for oversaturated conditions at fixed-time traffic signals. *Transp Res Record* 1091:48–59
- Dafermos S (1980) Traffic equilibrium and variational inequalities. *Transp Sci* 14(1):42–54
- Daganzo CF, Laval JA (2005) On the numerical treatment of moving bottlenecks. *Transp Res Part B Methodol* 39(1):31–46
- Daniel J, Fambro DB et al (1996) Accounting for non-random arrivals in estimate of delay at signalized intersections. *Transp Res Rec* 1555:9–16
- Dantzig GB (1957) The shortest route problem. *Oper Res* 5:270–273
- Dial R (1971) A probabilistic multipath traffic assignment model which obviates path enumeration. *Transp Res* 5:83–111
- Dijkstra EW (1959) A note on two problems in connection with graphics. *Numer Math* 1:209–271
- Dion F, Rakha H (2006) Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transp Res Part B* 40:745–766
- Dion F, Rakha H et al (2004a) Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transp Res Part B Methodol* 38(2):99–122
- Dion F, Rakha H et al (2004b) Evaluation of potential transit signal priority benefits along a fixed-time signalized arterial. *J Transp Eng* 130(3):294–303

- Elefteriadou L, Fang C et al (2005) Methodology for evaluating the operational performance of interchange ramp terminals. *Transp Res Rec* 1920:13–24
- Engelbrecht RJ, Fambro DB et al (1996) Validation of generalized delay model for oversaturated conditions. *Transp Res Rec* 1572:122–130
- Evans JL, Elefteriadou L et al (2001) Probability of breakdown at freeway merges using Markov chains. *Transp Res Part B Methodol* 35(3):237–254
- Fambro DB, Roupail NM (1996) Generalized delay model for signalized intersections and arterial streets. *Transp Res Rec* 1572:112–121
- Fang FC, Elefteriadou L et al (2003) Using fuzzy clustering of user perception to define levels of service at signalized intersections. *J Transp Eng* 129(6):657–663
- Fisk C (1979) More paradoxes in the equilibrium assignment problem. *Transp Res* 13B:305–309
- Flannery A, Kharoufeh JP et al (2005) Queuing delay models for single-lane roundabouts. *Civ Eng Environ Syst* 22(3):133–150
- Frank M (1981) The Braess paradox. *Math Program* 20:283–302
- Frank M, Wolfe P (1956) An algorithm of quadratic programming. *Nav Res Logist* 3:95–110
- Friesz TL, Luque J, Tobin RL, Wie B-W (1989) Dynamic network traffic assignment considered as a continuous time optimal control problem. *Oper Res* 37(6):893–901
- Friesz TL, Bernstein D, Smith TE, Tobin RL, Wie BW (1993) A variational inequality formulation of the dynamic network user equilibrium problem. *Oper Res* 41(1):179–191
- Ghali MO, Smith MJ (1995) A model for the dynamic system optimum traffic assignment problem. *Transp Res* 29B(3):155–170
- Greenshields BD (1934) A study of traffic capacity. *Proc Highway Res Board* 14:448–477
- Hagring O, Roupail NM et al (2003) Comparison of capacity models for two-lane roundabouts. *Transp Res Rec* 1852:114–123
- Hall MD, Van Vliet D et al (1980) SATURN a simulation assignment model of the evaluation of traffic management schemes. *Traffic Eng Control* 4:167–176
- Hawas YE (1995) A decentralized architecture and local search procedures for real-time route guidance in congested vehicular traffic networks. University of Texas, Austin
- Hawas YE (2004) Development and calibration of route choice utility models: neuro-fuzzy approach. *J Transp Eng* 130(2):171–182
- Hawas YE, Mahmassani HS (1995) A decentralized scheme for real-time route guidance in vehicular traffic networks. In: *Second world congress on intelligent transport systems*, Yokohama, 1995, pp 1965–1963
- Hawas YE, Mahmassani HS (1997) Comparative analysis of robustness of centralized and distributed network route control systems in incident situations. *Transp Res Rec* 1537:83–90
- Hawas YE, Mahmassani HS, Chang GL, Taylor R, Peeta S, Ziliaskopoulos A (1997) Development of dynasmart-X software for real-time dynamic traffic assignment. Center for Transportation Research, The University of Texas, Austin
- Hellinga BR, Van Aerde M (1998) Estimating dynamic O-D demands for a freeway corridor using loop detector data. Canadian Society for Civil Engineering, Halifax
- Ho JK (1980) A successive linear optimization approach to the dynamic traffic assignment problem. *Transp Sci* 14(4):295–305
- Hu SR, Madanat SM, Krogmeier JV, Peeta S (2001) Estimation of dynamic assignment matrices and OD demands using adaptive Kalman filtering. *Intell Transp Syst J* 6:281–300
- Chen H-K, Hsueh C-F (1998) A model and an algorithm for the dynamic user-optimal route choice problem. *Transp Res Part B Methodol* 32B(3):219–234
- Ishak S, Al-Deek H (2003) Performance evaluation of a short-term freeway traffic prediction model. Transportation Research Board 82nd annual meeting, Washington, DC
- Janson BN (1991a) Convergent algorithm for dynamic traffic assignment. *Transp Res Rec* 1328:69–80
- Janson BN (1991b) Dynamic traffic assignment for urban road networks. *Transp Res Part B Methodol* 25B:2–3
- Jayakrishnan R, Mahmassani HS (1990) Dynamic simulation assignment methodology to evaluate in-vehicle information strategies in urban traffic networks. In: *Winter simulation conference proceedings*, New Orleans 1990. 90 Winter simulation conference winter simulation conference proceedings. IEEE, Piscataway (IEEE cat n 90CH2926-4)
- Jayakrishnan R, Mahmassani HS (1991) Dynamic modeling framework of real-time guidance systems in general urban traffic networks. In: *Proceedings of the 2nd international conference on applications of advanced technologies in transportation engineering*, Minneapolis. ASCE, New York
- Jayakrishnan R, Mahmassani HS et al (1993) User-friendly simulation model for traffic networks with ATIS/ATMS. In: *Proceedings of the 5th international conference on computing in civil and building engineering V ICCCB E*, Anaheim 1993. ASCE, New York
- Jeffery DJ (1981) The potential benefits of route guidance. TRRL, Department of Transportation, Crowthorne
- Jha M, Madanat S, Peeta S (1998) Perception updating and day-to-day travel choice dynamics in traffic networks with information provision. *Transp Res Part C Emerg Technol* 6C(3):189–212
- Katsikopoulos KV, Duse-Anthony Y et al (2000) The framing of drivers' route choices when travel time information is provided under varying degrees of cognitive load. *J Hum Factors Ergon Soc* 42(3):470–481
- Kerner BS (2004a) The physics of traffic. Springer, Berlin
- Kerner BS (2004b) Three-phase traffic theory and highway capacity. *Physica A* 333(1–4):379–440
- Kerner BS (2005) Control of spatiotemporal congested traffic patterns at highway bottlenecks. *Physica A* 355(2–4):565–601
- Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase

- traffic theory: congestion nucleation in spatially non-homogeneous traffic. *Physica A* 364:473–492
- Kerner BS, Rehborn H et al (2004) Recognition and tracking of spatial-temporal congested traffic patterns on freeways. *Transp Res Part C Emerg Technol* 12(5):369–400
- Khattak AJ, Schofer JL, Koppelman FS (1993) Commuters' enroute diversion and return decisions: analysis and implications for advanced traveler information systems. *Transp Res Policy Pract* 27A(2):101
- Kim H, Baek S et al (2001) Origin-destination matrices estimated with a genetic algorithm from link traffic counts. *Transp Res Rec* 1771:156–163
- Koutsopoulos HN, Polydoropoulou A et al (1995) Travel simulators for data collection on driver behavior in the presence of information. *Transp Res Part C Emerg Technol* 3(3):143
- Krishnamurthy S, Coifman B (2004) Measuring freeway travel times using existing detector infrastructure. In: *Proceedings 7th international IEEE conference on intelligent transportation systems, ITSC*, Washington, DC
- Laval JA, Daganzo CF (2006) Lane-changing in traffic streams. *Transp Res Part B Methodol* 40(3):251–264
- Lawson TW, Lovell DJ et al (1996) Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck. *Transp Res Rec* 1572:140–147
- LeBlanc LJ (1975) An algorithm for discrete network design problem. *Transp Sci* 9:183–199
- LeBlanc LJ, Abdulaal M (1970) A comparison of user-optimum versus system-optimum traffic assignment in transportation network design. *Transp Res* 18B:115–121
- LeBlanc LJ, Morlok EK et al (1974) An accurate and efficient approach to equilibrium traffic assignment on congested networks. *Transp Res Rec* 491:12–23
- Lee S, Fambro D (1999) Application of the subset ARIMA model for short-term freeway traffic volume forecasting. *Transp Res Rec* 1678:179–188
- Leonard DR, Tough JB et al (1978) CONTRAM a traffic assignment model for predicting flows and queues during peak periods. TRRL SR 568. Transport Research Laboratory, Crowthorne
- Lertworawanich P, Elefteriadou L (2001) Capacity estimations for type B weaving areas based on gap acceptance. *Transp Res Rec* 1776:24–34
- Lertworawanich P, Elefteriadou L (2003) A methodology for estimating capacity at ramp weaves based on gap acceptance and linear optimization. *Transp Res Part B Methodol* 37(5):459–483
- Li Y (2001) Development of dynamic traffic assignment models for planning applications. Northwestern University, Evanston
- Li J, Roupail NM et al (1994) Overflow delay estimation for a simple intersection with fully actuated signal control. *Transp Res Rec* 1457:73–81
- Lighthill MJ, Witham GB (1955) On kinematic waves. I: Flood movement in long rivers, II A theory of traffic flow on long crowded roads. *Proc R Soc Lond A* 229:281–345
- Lorenz MR, Elefteriadou L (2001) Defining freeway capacity as function of breakdown probability. *Transp Res Rec* 1776:43–51
- Lotan T (1997) Effects of familiarity on route choice behavior in the presence of information. *Transp Res Part C Emerg Technol* 5(3–4):225–243
- Mahmassani H, Jou R-C (2000) Transferring insights into commuter behavior dynamics from laboratory experiments to field surveys. *Transp Res Part A Policy Pract* 34A(4):243–260
- Mahmassani H, Peeta S (1992) System optimal dynamic assignment for electronic route guidance in a congested traffic network. In: Gartner NH, Improta G (eds) *Urban traffic networks. Dynamic flow modelling and control*. Springer, Berlin, pp 3–37
- Mahmassani HS, Peeta S (1993) Network performance under system optimal and user equilibrium dynamic assignments: implications for ATIS. *Transp Res Rec* 1408:83–93
- Mahmassani HS, Peeta S (1995) System optimal dynamic assignment for electronic route guidance in a congested traffic network. In: Gartner NH, Improta G (eds) *Urban traffic networks: dynamic flow modeling and control*. Springer, Berlin, pp 3–37
- Mahmassani HS, Peeta S, Hu T, Ziliaskopoulos A (1993) Algorithm for dynamic route guidance in congested networks with multiple user information availability groups. In: *26th international symposium on automotive technology and automation*, Aachen
- Mahmassani HS, Chiu Y-C, Chang GL, Peeta S, Ziliaskopoulos A (1998a) Off-line laboratory test results for the DYNASMARTX real-time dynamic traffic assignment system. Center for Transportation Research, The University of Texas, Austin
- Mahmassani HS, Hawas Y, Abdelghany K, Abdelfatah A, Chiu Y-C, Kang Y, Chang GL, Peeta S, Taylor R, Ziliaskopoulos A (1998b) DYNASMART-X, vol II: Analytical and algorithmic aspects. Center for Transportation Research, The University of Texas, Austin
- Mahmassani HS, Hawas Y, Hu T-Y, Ziliaskopoulos A, Chang G-L, Peeta S, Taylor R (1998c) Development of Dynasmart-X software for real-time dynamic traffic assignment. Technical report ST067-85-Tast E (revised) submitted to Oak Ridge National Laboratory under subcontract 85X-SU565C
- Matsoukis EC (1986) Road traffic assignment, a review. Part I: Non-equilibrium methods. *Transp Plan Technol* 11:69–79
- Matsoukis EC, Michalopolos PC (1986) Road traffic assignment, a review. Part II: Equilibrium methods. *Transp Plan Technol* 11:117–135
- Mekky A (1995) Toll revenue and traffic study of highway 407 in Toronto. *Transp Res Rec* 1498:5–15
- Mekky A (1996) Modeling toll pricing strategies in greater Toronto areas. *Transp Res Rec* 1558:46–54
- Mekky A (1998) Evaluation of two tolling strategies for highway 407 in Toronto. *Transp Res Rec* 1649:17–25
- Merchant DK, Nemhauser GL (1978a) A model and an algorithm for the dynamic traffic assignment problems. *Transp Sci* 12(3):183–199

- Merchant DK, Nemhauser GL (1978b) Optimality conditions for a dynamic traffic assignment model. *Transp Sci* 12(3):200–207
- Minderhoud MM, Elefteriadou L (2003) Freeway weaving: comparison of highway capacity manual 2000 and Dutch guidelines. *Transp Res Rec* 1852:10–18
- Moskowitz K (1956) California method for assigning directed traffic to proposed freeways. *Bull Highw Res Board* 130:1–26
- Munnich LW Jr, Hubert HH et al (2007) L-394 MnPASS high-occupancy toll lanes planning and operational issues and outcomes (lessons learning in year 1). *Transp Res Rec* 1996:49–57
- Murchland JD (1970) Braess's paradox of traffic flow. *Transp Res* 4:391–394
- Nagel K (1996) Particle hopping model and traffic flow theory. *Phys Rev E* 53(5):4655–4672
- Nagel K, Schreckenberg M (1992) Cellular automaton model for freeway traffic. *J Phys* 2(20):2212–2229
- Nagel K, Schreckenberg M (1995) Traffic jam dynamics in stochastic cellular automata. *US D Energy, Los Alamos National Laboratory, LA-UR-95-2132*, Los Alamos
- Nakayama S, Kitamura R (2000) Route choice model with inductive learning. *Transp Res Rec* 1725:63–70
- Nakayama S, Kitamura R et al (2001) Drivers' route choice rules and network behavior: do drivers become rational and homogeneous through learning? *Transp Res Rec* 1752:62–68
- Newell GF (1965) Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Rev* 7:223–240
- Newell GF (1999) Delays caused by a queue at a freeway exit ramp. *Transp Res Part B Methodol* 33(5):337–350
- Nguyen S (1969) An algorithm for the assignment problem. *Transp Sci* 8:203–216
- Nie Y, Zhang HM et al (2005) Inferring origin-destination trip matrices with a decoupled GLS path flow estimator. *Transp Res Part B Methodol* 39(6):497–518
- Noonan J, Shearer O (1998) Intelligent transportation systems field operational test: cross-cutting study advance traveler information systems. *US Department of Transportation, Federal Highways Administration, Intelligent Transportation System, Washington, DC*
- Okutani I (1987) The Kalman filtering approaches in some transportation and traffic problems. In: *Proceedings of the tenth international symposium on transportation and traffic theory*. Elsevier, New York
- Park B (2002) Hybrid neuro-fuzzy application in short-term freeway traffic volume forecasting. *Transp Res Rec* 1802:190–196
- Park S, Rakha H (2006) Energy and environmental impacts of roadway grades. *Transp Res Rec* 1987:148–160
- Park D, Rilett LR (1998) Forecasting multiple-period freeway link travel times using modular neural networks. *Transp Res Rec* 1617:163–170
- Park D, Rilett LR (1999) Forecasting freeway link travel times with a multilayer feedforward neural network. *Comput-Aided Civ Infrastruct Eng* 14(5):357–367
- Park D, Rilett LR et al (1998) Forecasting multiple-period freeway link travel times using neural networks with expanded input nodes. In: *Proceedings of the 1998 5th international conference on applications of advanced technologies in transportation*, Newport Beach and *Proceedings of the international conference on applications of advanced technologies in transportation engineering 1998*, ASCE, Reston
- Park D, Rilett LR et al (1999) Spectral basis neural networks for real-time travel time forecasting. *J Transp Eng* 125(6):515–523
- Pavlis Y, Papageorgiou M (1999) Simple decentralized feedback strategies for route guidance in traffic networks. *Transp Sci* 33(3):264–278
- Peeta S (1994) System optimal dynamic traffic assignment in congested networks with advanced information systems. *University of Texas, Austin*
- Peeta S, Bulusu S (1999) Generalized singular value decomposition approach for consistent on-line dynamic traffic assignment. *Transp Res Rec* 1667:77
- Peeta S, Mahmassani HS (1995a) Multiple user classes real-time traffic assignment for online operations: a rolling horizon solution framework. *Transp Res Part C Emerg Technol* 3C(2):83
- Peeta S, Mahmassani HS (1995b) System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Ann Oper Res* 60:81–113
- Peeta S, Paz A (2006) Behavior-consistent within-day traffic routing under information provision. In: *IEEE intelligent transportation systems conference*, Toronto, pp 212–217
- Peeta S, Ramos JL (2006) Driver response to variable message signs-based traffic information. *Intell Transp Syst* 15(1):2–10
- Peeta S, Yang T-H (2000) Stability of large-scale dynamic traffic networks under on-line control strategies. In: *6th international conference on applications of advanced technologies in transportation engineering*, Singapore, paper no. 11 (eProceedings on CD), p 9
- Peeta S, Yang T-H (2003) Stability issues for dynamic traffic assignment. *Automatica* 39(1):21–34
- Peeta S, Yu JW (2004) Adaptability of a hybrid route choice model to incorporating driver behavior dynamics under information provision. *IEEE Trans Syst Man Cybern Part A Syst Humans* 34(2):243–256
- Peeta S, Yu JW (2006) Behavior-based consistency-seeking models as deployment alternatives to dynamic traffic assignment models. *Transp Res Part C Emerg Technol* 14(2):114–138
- Peeta S, Zhou C (1999a) On-line dynamic update heuristics for robust guidance. In: *International conference modeling and management in transportation*, Cracow, October 1999
- Peeta S, Zhou C (1999b) Robustness of the off-line a priori stochastic dynamic traffic assignment solution for on-line operations. *Transp Res Part C Emerg Technol* 7C(5):281–303
- Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. *Netw Spat Econ* 1(3–4):233
- Peeta S, Mahmassani HS et al (1991) Effectiveness of real-time information strategies in situations of non-recurrent congestion. In: *Proceedings of the 2nd international conference on applications of advanced*

- technologies in transportation engineering, Minneapolis. ASCE, New York
- Peeta S, Ramos JL, Pasupathy R (2000) Content of variable message signs and on-line driver behavior. *Transp Res Rec* 1725:102–108
- Rakha H (1990) An evaluation of the benefits of user and system optimised route guidance strategies. Civil Engineering, Queen's University, Kingston
- Rakha H, Ahn K (2004) Integration modeling framework for estimating mobile source emissions. *J Transp Eng* 130(2):183–193
- Rakha H, Arafeh M (2007) Tool for calibrating steady-state traffic stream and car-following models. In: Transportation research board annual meeting, Washington, DC, 22–25 Jan 2008
- Rakha H, Crowther B (2002) Comparison of greenshields, pipes, and van aerde car-following and traffic stream models. *Transp Res Rec* 1802:248–262
- Rakha H, Lucic I (2002) Variable power vehicle dynamics model for estimating maximum truck acceleration levels. *J Transp Eng* 128(5):412–419
- Rakha HA, Van Aerde MW (1996) Comparison of simulation modules of TRANSYT and integration models. *Transp Res Rec* 1566:1–7
- Rakha H, Zhang Y (2004a) INTEGRATION 2.30 framework for modeling lane-changing behavior in weaving sections. *Transp Res Rec* 1883:140–149
- Rakha H, Zhang Y (2004b) Sensitivity analysis of transit signal priority impacts on operation of a signalized intersection. *J Transp Eng* 130(6):796–804
- Rakha H, Van Aerde M et al (1989) Evaluating the benefits and interactions of route guidance and traffic control strategies using simulation. In: First vehicle navigation and information systems conference VNIS '89, Toronto. IEEE, Piscataway
- Rakha H, Van Aerde M et al (1998) Construction and calibration of a large-scale microsimulation model of the Salt Lake area. *Transp Res Rec* 1644:93–102
- Rakha H, Medina A et al (2000) Traffic signal coordination across jurisdictional boundaries: field evaluation of efficiency, energy, environmental, and safety impacts. *Transp Res Rec* 1727:42–51
- Rakha H, Kang Y-S et al (2001a) Estimating vehicle stops at undersaturated and oversaturated fixed-time signalized intersections. *Transp Res Rec* 1776:128–137
- Rakha H, Lucic I et al (2001b) Vehicle dynamics model for predicting maximum truck acceleration levels. *J Transp Eng* 127(5):418–425
- Rakha H, Ahn K et al (2004a) Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions. *Transp Res Part D Transp Environ* 9(1):49–74
- Rakha H, Pasumarthy P et al (2004b) Modeling longitudinal vehicle motion: issues and proposed solutions. In: Transport science and technology congress, Athens, Sep 2004
- Rakha H, Pasumarthy P et al (2004c) The INTEGRATION framework for modeling longitudinal vehicle motion. TRANSTEC, Athens
- Rakha H, Snare M et al (2004d) Vehicle dynamics model for estimating maximum light-duty vehicle acceleration levels. *Transp Res Rec* 1883:40–49
- Rakha H, Flintsch AM et al (2005a) Evaluating alternative truck management strategies along interstate 81. *Transp Res Rec* 1925:76–86
- Rakha H, Paramahamsan H et al (2005b) Comparison of static maximum likelihood origin-destination formulations. Transportation and traffic theory: flow, dynamics and human interaction. In: Proceedings of the 16th international symposium on transportation and traffic theory (ISTTT16), pp 693–716
- Ran B, Boyce DE (1996) A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. *Res Part C Emerg Technol* 4C(1):1–12
- Ran B, Shimazaki T (1989a) A general model and algorithm for the dynamic traffic assignment problems. In: Fifth world conference on transport research, transport policy, management and technology towards, Yokohama, 2001
- Ran B, Shimazaki T (1989b) Dynamic user equilibrium traffic assignment for congested transportation networks. In: Fifth world conference on transport research, Yokohama, 1989
- Ran B, Boyce DE, LeBlanc LJ (1993) A new class of instantaneous dynamic user-optimal traffic assignment models. *Oper Res* 41(1):192–202
- Ran B, Hall RW, Boyce DE (1996) A link-based variational inequality model for dynamic departure time/route choice. *Transp Res Methodol* 30B(1):31–46
- Randle J (1979) A convergence probabilistic road assignment model. *Traffic Eng Control* 11:519–521
- Richards PI (1956) Shock waves on the highway. *Oper Res* 4:42–51
- Rilett L, Aerde V (1993) Modeling route guidance using the integration model. In: Proceedings of the pacific rim trans tech conference, Seattle, 1993 and Proceedings of the ASCE international conference on applications of advanced technologies in transportation engineering. ASCE, New York
- Rilett L, Van Aerde M (1991a) Routing based on anticipated travel times. In: Proceedings of the 2nd international conference on applications of advanced technologies in transportation engineering, Minneapolis. ASCE, New York
- Rilett LR, van Aerde MW (1991b) Modelling distributed realtime route guidance strategies in a traffic network that exhibits the Braess paradox. In: Vehicle navigation & information systems conference proceedings part 2 (of 2). Dearborn, 1991. Proceedings society of automotive engineers n P-253. SAE, Warrendale
- Rilett LR, Van Aerde M et al (1991) Simulating the TravTek route guidance logic using the integration traffic model. In: Vehicle navigation & information systems conference proceedings part 2 (of 2). Dearborn, 1991. In: Proceedings society of automotive engineers n P-253, SAE. Warrendale
- Roughail NM (1988) Delay models for mixed platoon and secondary flows. *J Transp Eng* 114(2):131–152
- Roughail NM, Akcelik R (1992) Preliminary model of queue interaction at signalised paired intersections. In: Proceedings of the 16th ARRB conference, Perth, 9–12 November 1992. Congestion management proceedings conference of the Australian Road Research Board. Australian Road Research Board, Nunawading

- Schofer AJKFSKJL (1993) Stated preferences for investigating commuters' diversion propensity. *Transportation* 20(2):107–127
- Sheffi Y (1985) *Urban transportation networks: equilibrium analysis with mathematical programming methods*. Prentice Hall, Englewood Cliffs
- Sheffi Y, Powell W (1981) A comparison of stochastic and deterministic traffic assignment over congested networks. *Transp Res* 15B:65–88
- Shen W, Nie Y et al (2006) Path-based system optimal dynamic traffic assignment models: formulations and solution methods. In: *IEEE intelligent transportation systems conference IEEE*, Toronto, pp 1298–1303
- Sherali HD, Arora N, Hobeika AG (1997) Parameter optimization methods for estimating dynamic origin-destination triptables. *Transp Res Part B Methodol* 31B(2):141–157
- Sherali HD, Desai J et al (2006) A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. *Transp Res Part B* 40:857–871
- Simon HA (1947) Administrative behavior. *Am Polit Sci Rev* 41(6)
- Simon HA (1955) A behavioral model of rational choice. *Q J Econ* 69(1):99–118
- Simon H (1957) Models of man, social and rational. *Adm Sci Q* 2(2)
- Sivanandan R, Dion F et al (2003) Effect of variable-message signs in reducing railroad crossing impacts. *Transp Res Rec* 1844:85–93
- Smock R (1962) An iterative assignment approach to capacity-restraint on arterial networks. *Bull Highw Res Board* 347:226–257
- Srinivasan KK, Mahmassani HS (2000) Modeling inertia and compliance mechanisms in route choice behavior under realtime information. *Transp Res Rec* 1725:45–53
- Steinberg R, Zangwill WI (1983) The prevalence of braess' paradox. *Transp Sci* 17:301–318
- Stewart N (1980) Equilibrium versus system-optimal flow: some examples. *Transp Res* 14A:81–84
- Talaat H, Abdulhai B (2006) Modeling driver psychological deliberation during dynamic route selection processes. In: *2006 IEEE intelligent transportation systems conference*, Toronto, pp 695–700
- Tarko A, Roupail N et al (1993) Overflow delay at a signalized intersection approach influenced by an upstream signal. An analytical investigation. *Transp Res Rec* 1398:82–89
- Van Aerde M (1985) *Modelling of traffic flows, assignment and queueing in integrated freeway/traffic signal networks*. PhD, University of Waterloo, Waterloo
- Van Aerde M, Rakha H (1989) Development and potential of system optimized route guidance strategies. In: *IEEE vehicle navigation and information systems conference IEEE*, Toronto, pp 304–309
- Van Aerde M, Rakha H (2007) *INTEGRATION © Release 2.30 for Windows: user's guide vol I: fundamental model features*. M Van Aerde & Assoc, Ltd, Blacksburg
- Van Aerde M, Yagar S (1988a) Dynamic integrated freeway/traffic signal networks: a routeing-based modeling approach. *Transp Res* 22A(6):445–453
- Van Aerde M, Yagar S (1988b) Dynamic integrated freeway/traffic signal networks: problems and proposed solutions. *Transp Res* 22A(6):435–443
- Van Aerde, M, Hellinga BR et al (1993) QUEENSOD: a method for estimating time varying origin-destination demands for freeway corridors/networks. In: *72nd annual meeting of the transportation research board*, Washington, DC
- Van Aerde M, Rakha H et al (2003) Estimation of origin-destination matrices: relationship between practical and theoretical considerations. *Transp Res Rec* 1831:122–130
- Van Der Zijpp NJ, De Romph E (1997) A dynamic traffic forecasting application on the Amsterdam beltway. *Int J Forecast* 13:87–103
- Van Vliet D (1976) Road assignment. *Transp Res* 10:137–157
- Van Vliet D (1982) SATURN a modern assignment model. *Traffic Eng Control* 12:578–581
- Van Zuylen JH, Willumsen LG (1980) The most likely trip matrix estimated from traffic counts. *Transp Res* 14B:281–293
- Walker N, Fain WB et al (1997) Aging and decision making: driving-related problem solving. *J Hum Factors Ergon Soc* 39(3):438–444(7)
- Waller ST (2000) *Optimization and control of stochastic dynamic transportation systems: formulations, solution methodologies, and computational experience*. PhD, Northwestern University, Evanston
- Waller ST, Ziliaskopoulos AK (2006) A chance-constrained based stochastic dynamic traffic assignment model: analysis, formulation and solution algorithms. *Transp Res Part C Emerg Technol* 14(6):418–427
- Wardrop J (1952) Some theoretical aspects of road traffic research. *Institute of Civil Engineers*, pp 325–362
- Webster F (1958) *Traffic signal settings*. HMsSO Road Research Laboratory, London
- Webster FV, Cobbe BM (1966) *Traffic signals*. HMsSO Road Research Laboratory, London
- Wie BW (1991) Dynamic analysis of user-optimized network flows with elastic travel demand. *Transp Res Rec* 1328:81–87
- Willumsen LG (1978) Estimation of an O-D matrix from traffic counts: a review. *Institute for Transport Studies*, Working paper no 99, Leeds University, Leeds
- Wilson AG (1970) *Entropy in urban and regional modeling*. Pion, London
- Wu J, Chang G-L (1996) Estimation of time-varying origin-destination distributions with dynamic screenline flows. *Transp Res Part B Methodol* 30B(4):277–290
- Yagar S (1971) Dynamic traffic assignment by individual path minimization and queueing. *Transp Res* 5:179–196
- Yagar S (1974) Dynamic traffic assignment by individual path minimization and queueing. *Transp Res* 5:179–196

- Yagar S (1975) CORQ a model for predicting flows and queues in a road corridor. *Transp Res* 553:77–87
- Yagar S (1976) Measures of the sensitivity and effectiveness of the CORQ traffic model. *Transp Res Rec* 562:38–48
- Yang T-H (2001) Deployable stable traffic assignment models for control in dynamic traffic networks: a dynamical systems approach. PhD, Purdue University, West Lafayette
- Yang Q, Ben-Akiva ME (2000) Simulation laboratory for evaluating dynamic traffic management systems. *Transp Res Rec* 1710:122–130
- Zhou X, Mahmassani HS (2006) Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Trans Intell Transp Syst* 7(1):105–114
- Zhou Y, Sachse T (1997) A few practical problems on the application of OD-estimation in motorway networks. *TOP* 5(1):61–80
- Ziliaskopoulos AK (2000) A linear programming model for the single destination system optimum dynamic traffic assignment problem. *Transp Sci* 34(1):37–49
- Ziliaskopoulos AK, Waller ST (2000) An Internet-based geographic information system that integrates data, models and users for transportation applications. *Transp Res Part C Emerg Technol* 8C:1–6
- Ziliaskopoulos A, Wardell W (2000) Intermodal optimum path algorithm for multimodal networks with dynamic arc travel times and switching delays. *Eur J Oper Res* 125(3):486–502

Optimization and Control of Urban Traffic Networks

Nathan H. Gartner and Chronis Stamatiadis
University of Massachusetts, Lowell, MA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Traffic Signal Control
The Basic Model – Single Intersection
Arterial Systems – Progression Models
Delay-Based Models – Cyclic Flow Profiles
Demand-Responsive Models
Multi-Level Traffic Control Strategies
Control System Performance – Analysis of Complexity
Future Directions
Bibliography

Glossary

Assignment (traffic) The allocation of traffic volumes to routes.
Arterial A road with two or more intersections.
Bandwidth Time interval in an arterial progressive system during which vehicles can travel unimpeded.
Controller A device which controls the sequence and duration of indications displayed by traffic signals.
Coordination The establishment of a definite timing relationship between adjacent traffic signals.
Cycle time The time required for one complete sequence of signal indications.
Delay (traffic) The time lost by vehicle(s) due to traffic interference or control devices.

Detector (traffic) A device to detect the presence or passage of a vehicle in the roadway.
Flow (traffic) The rate at which vehicles cross a given line on the road; the number of vehicles per unit of time.
Headway The time interval between successive vehicle crossings of a given line on the road.
Intelligent transportation system The addition of information and communications technology to enhance performance of transport infrastructure and of vehicles.
Offset The time difference between the start of green at one intersection and the start of green at a subsequent intersection, or with respect to a system time base.
Phase The time interval of a cycle time allocated to any combination of movements receiving the right-of-way simultaneously.
Performance index/disutility index A numerical quantity used to measure the performance of a traffic in a network in an optimization or a simulation model; also objective function.
Platoon A tight group of vehicles traveling along an arterial.
Progression A timing plan enabling a platoon to travel unimpeded along an arterial.
Saturation flow The rate of flow at which vehicles are discharged from a queue stopped at a signal; the maximum rate of flow on a street.
Synchronization A condition under which traffic signals operate with the same cycle time.
Volume (traffic) The number of vehicles crossing a line on the road per hour; traffic flow.

Definition of the Subject

Much of our economic and social life is dependent on communication and transportation systems. Travel in urban networks represents an interaction between the demand for transportation and the supply of transportation means and facilities which includes the vehicles, the road networks,

as well as the control systems that govern them. The critical elements within these systems are junctions, or intersections, which are controlled by means of traffic signals. Public authorities, which are responsible for the operation of those systems, develop control policies that have a dominant impact on the quality of travel and the level of service provided by the networks. An understanding of the basic models involved in urban traffic control is essential for the development of optimal operating policies that would lead to the effective performance of urban traffic networks.

Introduction

The critical elements within an urban traffic network are the junctions, or the intersections, many of which are controlled by traffic signals. Therefore, traffic control signals are a key determinant for the efficient operation of urban street networks and an essential element of Intelligent Transportation Systems.

In this chapter we focus on the complex array of urban street networks and their control by means of traffic signals. We start by delineating the basic models involved in traffic flow at single intersections. We then present control models on arterial streets and networks, both progression models and delay-based models. Next we introduce demand-responsive systems and we conclude by outlining a strategy for multi-level control and an analysis of the complexity involved in the performance of the different systems.

Traffic Signal Control

Traffic control signals are defined as “power-operated traffic devices which alternately direct traffic to stop and to proceed.” More specifically, traffic signals are used to control the assignment of right-of-way at locations where conflicts exist or where passive devices, such as signs and markings, do not provide the necessary flexibility of control to properly move traffic in a safe and efficient manner.

Traffic control signals are usually described as either *pre-timed* or *traffic actuated*. Each type may be used in either an independent (isolated) or

interconnected (system) application. *Pre-timed control* assigns the right of way at an intersection according to a predetermined schedule. The length of the time interval for each signal indication in the cycle is fixed, based on historic traffic patterns. *Actuated control* differs from pre-timed in that signal phases are not of fixed length. Through the use of vehicle detectors, this type of control assigns the right of way on the basis of actual traffic conditions (demands) within given limitations. The full range of actuated control capabilities depends on the type of equipment employed and the operational requirements.

Traffic signal timings are divided into time phases. A signal phase may be defined as that part of the cycle length allocated to a traffic movement receiving the right of way or to any combination of traffic movements receiving the right of way simultaneously. A traffic movement is a single vehicular movement, a single pedestrian movement, or a combination of vehicular and pedestrian movements. The sum of all traffic phases at an intersection is equal to the cycle length.

The Basic Model – Single Intersection

The material in this section is based on models developed by F.V. Webster at the British Road Research Laboratory (Webster 1958) for estimating delays to vehicles at fixed-time traffic signals and for computing optimum control settings for such signals. The methods developed in this study can be applied both to fixed-time signals and to vehicle-actuation.

Webster’s classical results provide a more definite basis for setting fixed-time signals than any which have so far been published. They are else applicable to those vehicle actuated signals where the green periods owing to heavy traffic demands are frequently running to maximum, giving in effect fixed-time control. In addition, linked systems of traffic signals often work on a fixed cycle which has been set to meet the requirements of the main intersection of the system and the desired speed of traffic along the road.

Flow Model

The first step in the analysis of traffic control signals is to determine performance, most often measured

by the delay caused to traffic. Traffic may be considered as arriving at random provided that the point at which it is observed is some distance from a disturbing factor such as a controlled intersection. Total delay to traffic is computed by the area under the queue-length curve. The average delay is obtained by dividing the total delay by the number of arrivals. This is illustrated in Fig. 4 below.

When the signal is red, arriving traffic queues at the signal stop line. After a green period commences, a certain time elapses while vehicles are accelerating to normal running speed, but after a few seconds the queue discharges at a constant rate, called the saturation flow. This is illustrated in the time-space (or distance) diagram of Fig. 1. If there is still a queue at the end of the green period, some vehicles will make use of the yellow interval to cross the intersection. In these circumstances traffic moves on both green and yellow signals but the discharge rate is less than the saturation flow both at the beginning and at the end of the right-of-way period, as shown in Fig. 2. The green and yellow together ($k + a$) may be replaced by an 'effective' green (g) and a 'lost' time (l), such that the product of the effective green and the saturation flow is equal to the correct number of vehicles (say, b) discharged from the queue on the average in a saturated green period (i.e. a green period during which the queue never clears).

Thus, $k + a = g + 1$ and $b = g \cdot s$, where s is the saturation flow rate.

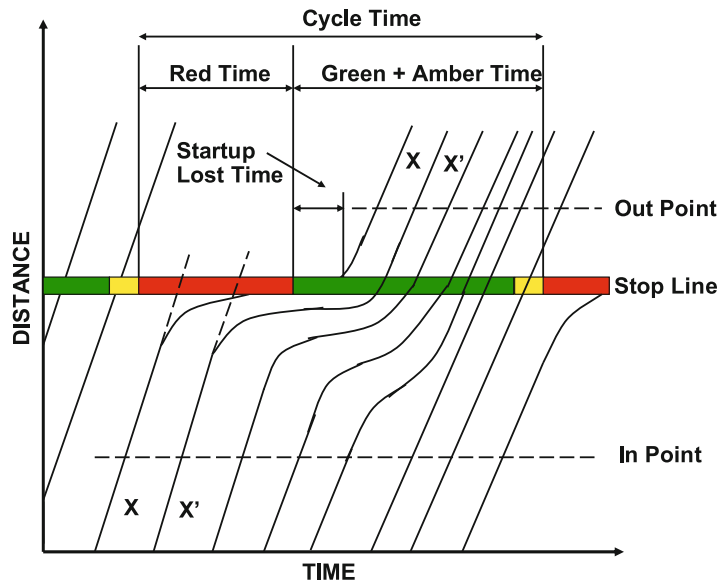
In the basic model presented in this section the saturation flow is assumed to be constant. The signal sequence on any approach is thus reduced to an 'effective' green period and a 'red' period which comprises all the times when traffic cannot run, i.e. red periods plus 'lost' time. The delay to traffic on a single approach to an intersection controlled by fixed-time traffic signals is computed over a range of values of green times, cycle times, traffic flows and saturation flows covering most practical conditions.

Signalized Intersections – Fluid Traffic Model

There are several models that may be used in the investigation of queues and delays at signalized intersections. In this section a continuum (i.e., deterministic) or fluid model is first considered for which basic measures of queue length and delay are developed (Gerlough and Huber 1975). In the next section the stochastic nature of traffic is taken into consideration. In estimating delay at intersections, traffic flow is considered to consist of identical passenger car units (pcu's). A truck, for example, may be considered as 1.5 or 2 equivalent pcu's and a turning vehicle may also be assigned some value depending on the type of maneuver that is made.

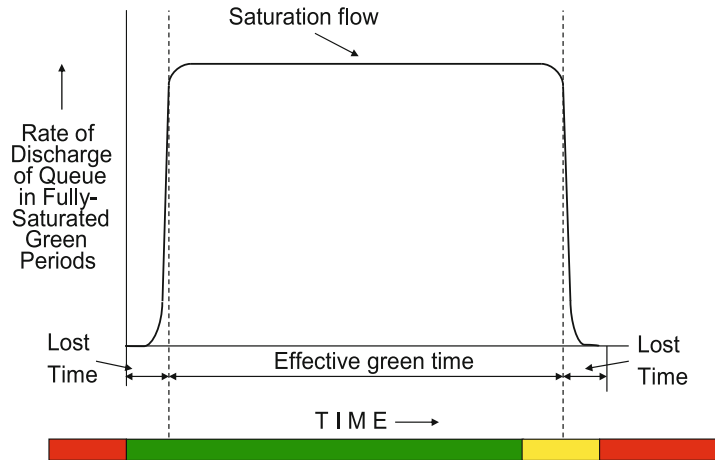
Optimization and Control of Urban Traffic Networks,

Fig. 1 Distance/time diagram for flow through intersection



Optimization and Control of Urban Traffic Networks,

Fig. 2 Variation of discharge rate of queue with time in a fully-saturated green period



The following notation is used:

- c = the cycle time (sec);
- g = the effective green time (sec);
- r = the effective red time (sec);
- q = the average arrival rate of traffic on the approach (pcu's/sec);
- s = the saturation flow on the approach (pcu's/sec);
- d = the average delay to a pcu's on the approach (sec);
- Q_0 = the overflow (pcu's);
- $\lambda = g/c$ (the proportion of the cycle that is effectively green);
- $y = q/s$ (the ratio of average arrival rate to saturation flow); and
- $x = q \cdot c/g \cdot s$ (the ratio of average number of arrivals/cycle to the maximum number of departures/cycle).

Thus, $r + g = c$ and $\lambda \cdot x = y$. The ratio x is called the degree of saturation of the approach and y is called the flow ratio of the approach.

The effective green time is the portion of the cycle time during which pcu's are assumed to pass the signal at a constant rate s , provided vehicles are waiting on the approach. Greenshields (Greenshields et al. 1947) observed that the times to cross the signal stop line are 3.8, 3.1, 2.7, 2.4, 2.2, 2.1, 2.1 s for the consecutive queued vehicles. The cumulative time for a queue of n stopped vehicles to pass a signal can then be given by:

$$\text{Cumulative time} = 14.2 + 2.1(n - 5) \text{ sec} \quad \text{for } n \geq 5.$$

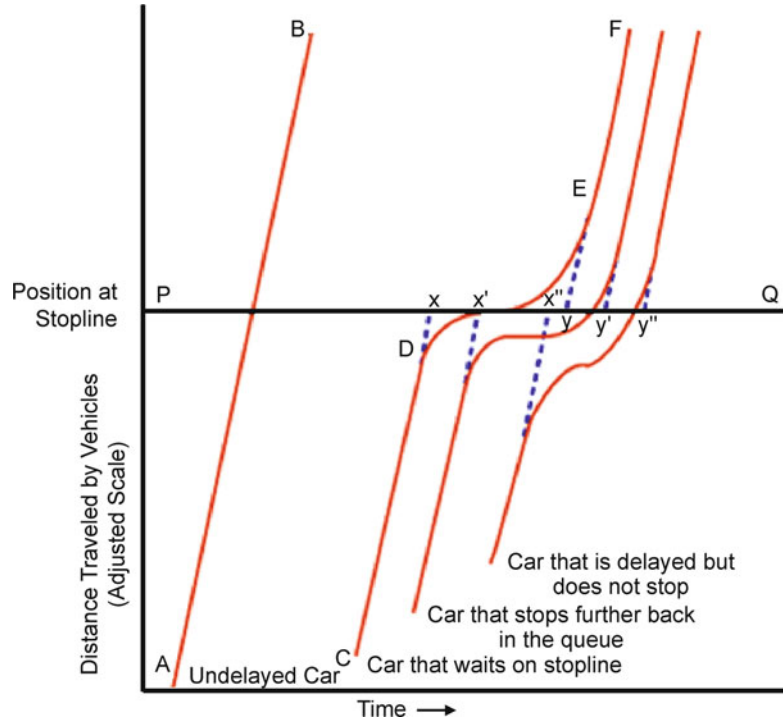
Had all the vehicles departed at the saturation rate $s = 1/2.1 \text{ veh/s} = 1714 \text{ veh/h}$, the first five vehicles would have required 10.5 s; that is, the effective green is the signal green time less 3.7 s. In most studies it is also assumed that a waiting queue of vehicles will take advantage of the yellow clearance interval, although the effective green time may be adjusted to reflect particular operating conditions.

Delay can be calculated by illustrating the meaning of arrival time and departure time for a pcu on an approach. They are demonstrated by reference to Fig. 3, in which distance-time curves are plotted for each of four vehicles. AB represents the passage of an un-delayed vehicle, where the line PQ represents the stopline at which the first vehicle waits when there is a queue. CDEF represents the trajectory of the first vehicle that is delayed by a signal. The straight portions of CD and EF are parallel to AB and projected to meet PQ and X and Y so that the length XY is the delay to the first vehicle. Similarly, X'Y' and X''Y'' represent the delays for the next two vehicles. XX' and X'X'' represent the arrival headways.

Continuum Model for a Pre-timed Signal

Representation of a continuum model of traffic flow at a signalized intersection is given in Fig. 4. The vertical axis represents the number of vehicle arrivals of the stop line and the horizontal

Optimization and Control of Urban Traffic Networks, Fig. 3 Vehicle arrival-departure times and delay of the stop-line



axis the time t . The figure illustrates the behavior when the capacity of the green interval exceeds the arrival during the green + red time i.e., an under-saturated period. The top figure illustrates the queue length as a function of time (t is the queue clearance time, measured from the start of green). The bottom figure illustrates the cumulative arrival and departure functions. In Fig. 4 the vertical distance ca represents the number of vehicles that have accumulated since the signal entered the red phase. The horizontal distance ab represents the total time from arrival to departure for any given vehicle. The shaded areas, in each figure, represent the total vehicle delay.

The following measures of queue behavior are developed:

1. Time after start of green that queue is dissipated (t_0);
2. Proportion of cycle with queue (P_q);
3. Proportion of vehicles stopped (P_s);
4. Maximum number of vehicles in queue (Q_m);
5. Average number of vehicles in queue ($\bar{Q}$);
6. Total vehicle-time of delay per cycle (D);
7. Average individual vehicle delay (d);

8. Maximum individual vehicular delay (d_m);

The formulas for these measures are developed from simple geometric relationships:

1. For any given cycle it is evident that at time t_0 after the start of green, the total number of arrivals must equal the number discharged: $q \cdot (r + t_0) = s \cdot t_0$. Letting $y = q/s$ we obtain:

$$t_0 = y \cdot r / (1 - y).$$

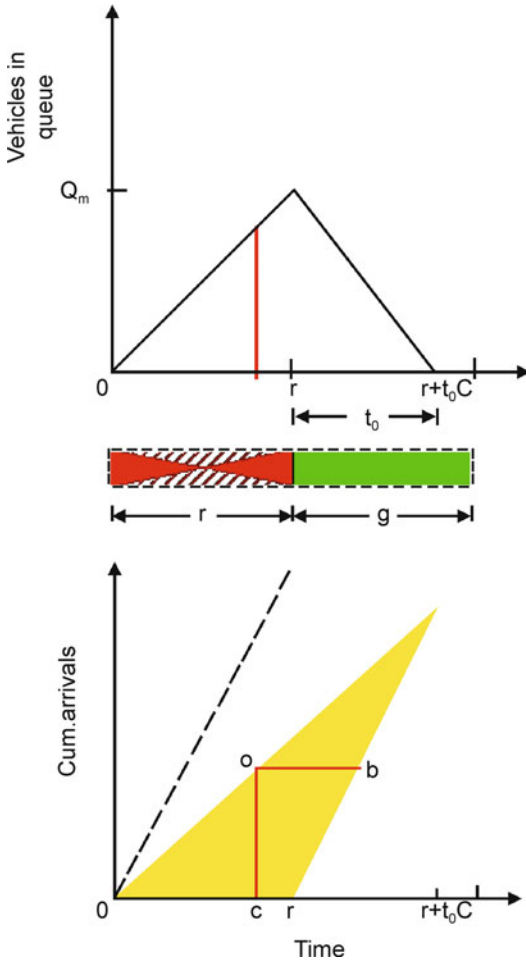
2. Proportion of vehicles stopped is equal to queue time/cycle length

$$P_q = (r + t_0) / c.$$

3. Proportion of vehicles stopped is equal to vehicles stopped/total vehicles per cycle

$$P_s \cdot q \cdot (r + t_0) / q \cdot (r + g) = t_0 / (y \cdot c).$$

4. The maximum number of vehicles in queue can be seen by inspection to be the height of the triangle at $t = r$, i.e., at the end of the red period.



Optimization and Control of Urban Traffic Networks,
Fig. 4 Cumulative arrival-departure diagrams and queuing at a signal (continuum model)

$$Q_m = q \cdot r.$$

5. The average number of vehicles in the queue, over the total length of cycle c is

$$\bar{Q} = \frac{(qr/2)r + (qr/2)t_0 + o(g - t_0)}{r + t_0 + g - t_0}.$$

which yields: $\bar{Q} = (r + t_0) \cdot q \cdot r / (2 \cdot c)$.

6. The total vehicle-time of delay during the cycle is given by the area of the triangle

$$D = (q \cdot r/2)(r + t_0) = (q \cdot r/2)(r/(1 - y)) = q \cdot r^2 / (2(1 - y)).$$

7. The average individual delay is given by dividing the total delay by the number of vehicles, or

$$d = \left[\frac{qr^2}{2(1 - y)} \right] \frac{1}{qc} = \frac{r^2}{2c(1 - y)}.$$

8. The maximum individual vehicular delay will be seen from Fig. 4 to be

$$d_m = r.$$

If the possible departures during the cycle sg (i.e., the capacity flow) are less than the arrivals qc , the queue grows with each successive cycle and the foregoing formulas are no longer applicable. We have, then, over-saturation.

Average Delay per Vehicle

Webster found that the results of his computations using a stochastic simulation model could be expressed to a close approximation by the equation

$$d = \frac{c(1 - \lambda)^2}{2(1 - \lambda x)} + \frac{x^2}{2q(1 - x)} - 0.65 \left(\frac{c}{q^2} \right)^{\frac{1}{3}} x^{(2+5\lambda)} \quad (1)$$

where:

d = average delay per vehicle on the particular arm of the intersection

c = cycle time

λ = proportion of the cycle which is effectively green for the phase under consideration (i.e. g/c)

q = flow

s = saturation flow

x = the degree of saturation – this is the ratio of the actual flow to the maximum flow which can be passed through the intersection from this arm, and is given by: $x = q/\lambda s$.

An estimate of the average delay per vehicle is required to evaluate the performance of traffic control signals. To enable the expected delay to

be estimated more easily, Eq. (1) can be rewritten as

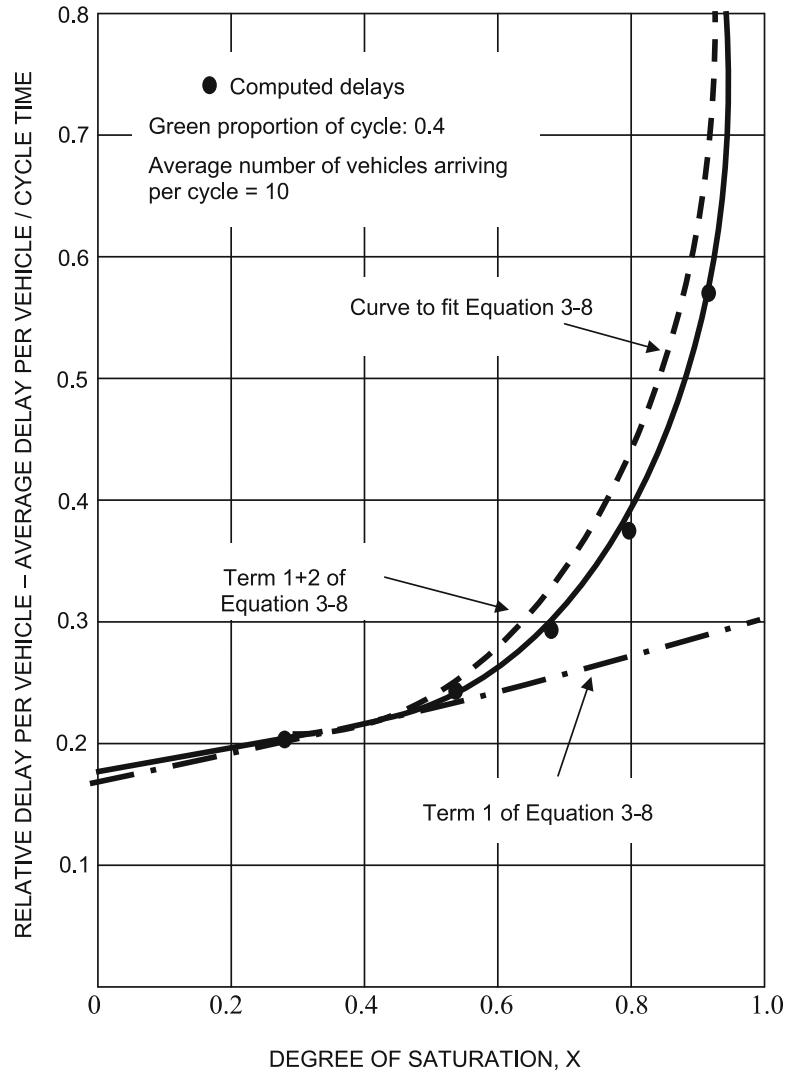
$$d = cA + \frac{B}{q} - C \quad (2)$$

where $A = (1 - \lambda)^2/2(1 - \lambda x)$, $B = x^2/2(1 - x)$ and C is a correction term.

The first two terms in Eq. (1) have a theoretical underpinning but the last term is purely empirical. The first term is the expression for the delay when traffic can be considered to be arriving at a uniform rate. It is identical to the formula in Item No. 7 above. Although the agreement between

computed delays and those derived from this term is fairly good at low flows, as shown by Fig. 5, it is not so at higher values where the computed delays, owing to the random nature of the arrivals, are far in excess of values calculated from this term only. The second term makes some allowance for the random nature of the arrivals. It is an expression for the delay experienced by vehicles arriving randomly in time at a 'bottle-neck', queuing up, and leaving at constant intervals. In queuing terminology it is the average delay of an M/D/1 queue. The addition of the second term improves the correspondence with Eq. (1), but overestimates it. To avoid the need

Optimization and Control of Urban Traffic Networks, Fig. 5 Typical fixed-time delay curve



to calculate the correction factor C , a suitable approximation can be obtained by one of the following:

$$d = \left[cA + \frac{B}{q} \right] \frac{100 - P}{100} \quad (3)$$

where (a) $P = 10\%$, or (b) $P = 15x$, since the value of the correction term C is generally in the range 5–15% of d .

The ‘lost time’ illustrated in Fig. 2 plays an important role in determining the efficiency of the traffic signal operation. Information available on lost time suggests a value of about 4.5 s per phase plus any all-red periods. This is composed of 2.5 s start-up lost time plus half the typical 4 s yellow time. Since lost time depends on gradients, type of traffic, etc. its value will vary from site to site and even at the same site will probably vary with time of day.

Optimum Settings of Fixed-Time Signals

In general, all approaches belonging to the same phase will have the same green period even though the traffic requirements of the approaches may be different. To simplify calculations, and with negligible loss of accuracy, each phase can be represented by one approach only – the one with the highest ratio of flow to saturation flow. Let this ratio be denoted by the symbol y . In deriving the optimum green times and cycle time the empirical correction term of the delay equation is neglected since calculations show that its variation with respect to these quantities is slight.

Green Times Least delay to traffic is obtained when the green periods of the phases are in proportion to the corresponding ratios of flow to saturation flow, assuming this ratio to be the same for all arms of the same phase. This division of the cycle time makes the capacity of the phases proportional to the average flows of the phases. That this is approximately the best division of the cycle time is shown by the examples given below.

The total delay experienced by vehicles at an intersection was calculated using the approximate form of the delay Eq. (2). The lost time was assumed to be 10 s per cycle and delays were calculated for a variety of cycle times and ratios of the effective green

times. A typical result is shown in Fig. 6, where the ratio of the y values is 2.0. It can be seen that the best ratios of the effective green times is between 1.88 and 2.17 over a range of cycle times of 35–80 s.

Cycle Time In deriving an expression for the optimum cycle time it is assumed that the effective green times of the phases are in the ratio of their respective y values. Equation (3) representing the delay to traffic on one arm of an intersection is modified to cover the general case of an intersection with n phases, and the modified equation is differentiated with respect to cycle time to determine the value of cycle time which gives the least delay of all traffic using the intersection.

The derivation of the optimum cycle-time formula is given in (Webster 1958). This formula is too complicated for most purposes and a simple approximation is therefore derived as follows:

$$c_0 = \frac{1.5L + 5}{1 - Y} \text{ sec} \quad (4)$$

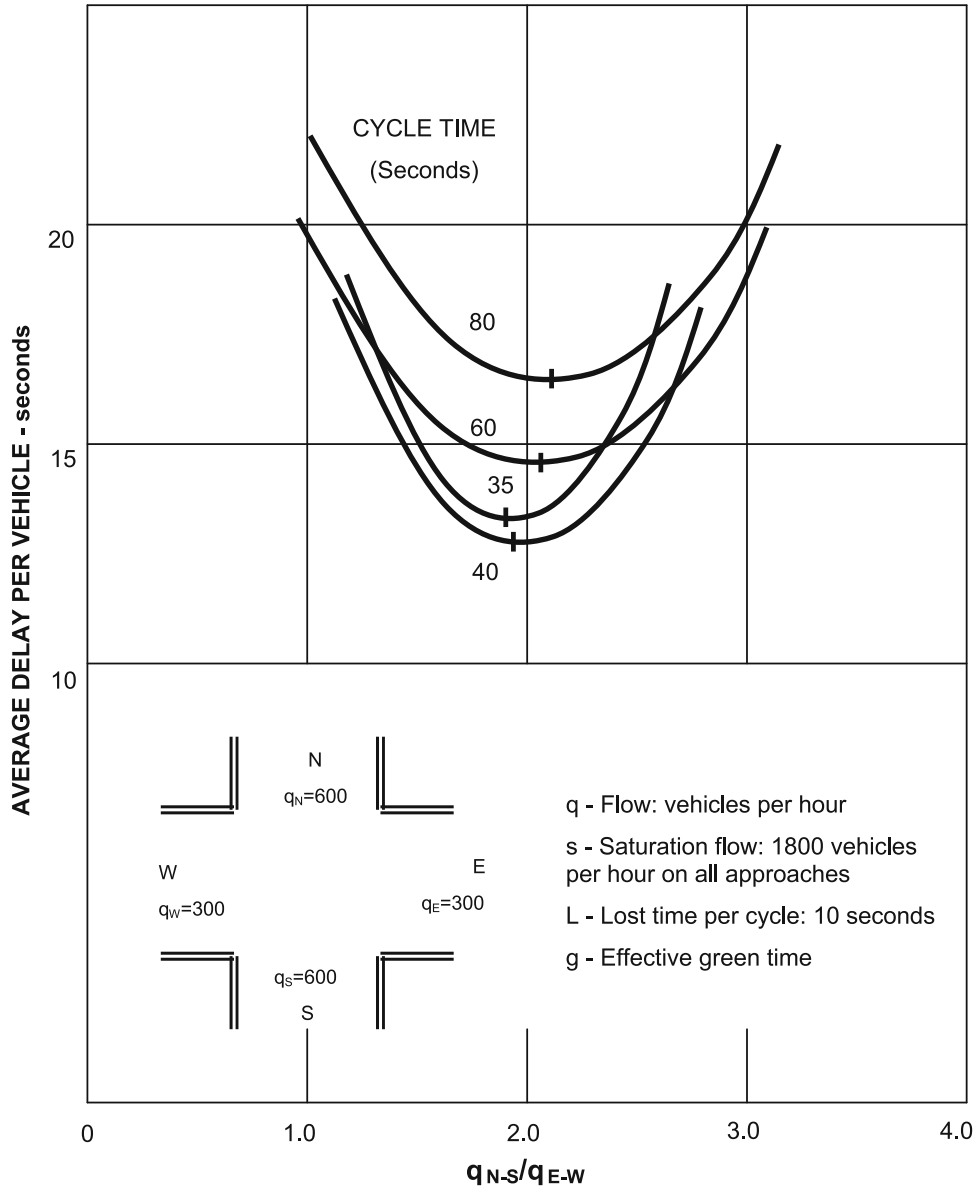
where Y is the sum of the y values and refers to the intersection as a whole and L is the total lost time per cycle in seconds. The lost time can be expressed by: $L = nl + R$.

Where:

- n is the number of phases
- l is the average lost time per phase (excluding any all-red periods)
- R is the time during each cycle when all signals display red (including red-with-amber) simultaneously.

The possible aspect scheduling for a 2-phase signal is shown in Fig. 7.

Several examples of hypothetical intersections (including some fairly extreme cases) have been studied to show the effect of changes in cycle time on the delay. For a symmetrical intersection, values of delay have been computed by Eq. (1) and are shown in Fig. 8 plotted against the cycle length for several values of the total flow entering the intersection. The cycle time giving the least delay in each case is approximately twice the least possible cycle which will just allow the traffic to pass through (although with long delays). The latter is called the minimum cycle, and is the

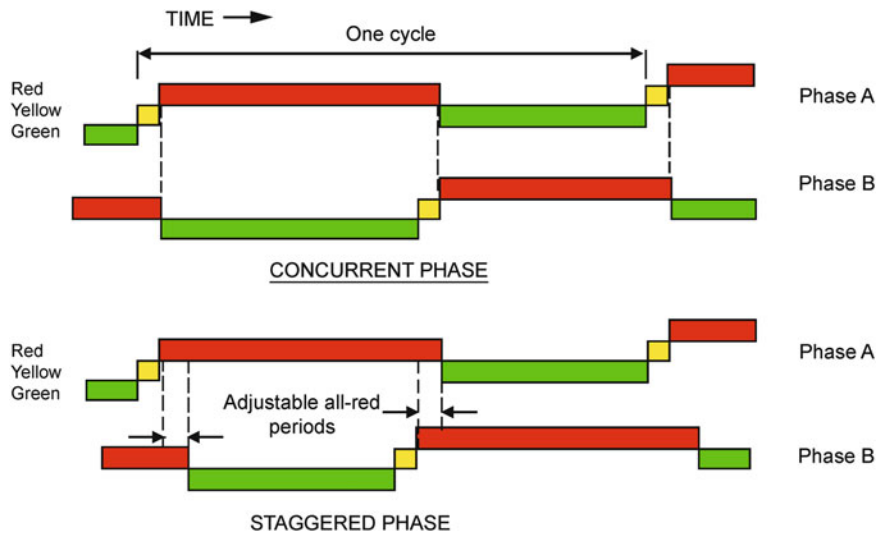


Optimization and Control of Urban Traffic Networks, Fig. 6 Effect on delay of the variation of the ratio of green periods

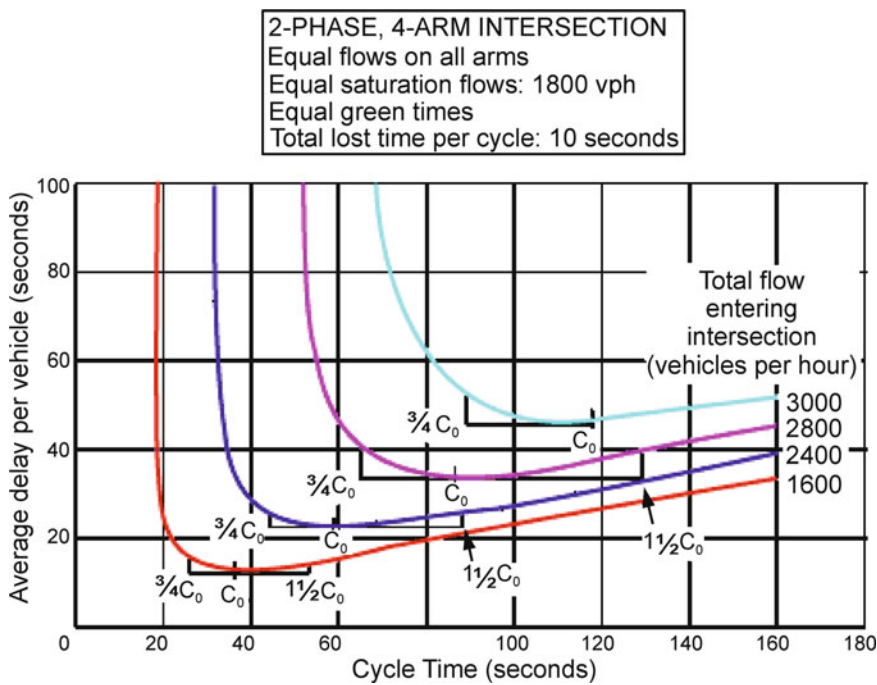
vertical asymptote to the delay/cycle-time curve. When traffic is of a truly random character the minimum cycle is associated with infinite delay. For uniform flow it is the cycle such that all traffic arriving during the cycle can just be discharged during the green. We can approximate the optimum cycle as follows:

$$c_{opt} \cong 2c_{min} = \frac{2L}{1 - Y}$$

From graphs such as Fig. 8, it was found that in most practical cases the delay for cycle times within the range $\frac{3}{4}$ to $1\frac{1}{2}$ times the optimum



Optimization and Control of Urban Traffic Networks, Fig. 7 Possible phase aspect scheduling at a two-phase intersection



Optimization and Control of Urban Traffic Networks, Fig. 8 Effect on delay of variation of the cycle length

value is never more than 10–20% greater than that given by the optimum cycle. This fact can be used in deducing a compromise cycle time when the level of flow varies considerably throughout the day. It would be better either to change the cycle

time to take account of this, or, as is more common, to use vehicle-actuated signals. However, for a single setting of fixed time signals the simple approximate method outlined below may be used.

1. Calculate the optimum cycle for each hour of the day when the traffic flow is medium or heavy, e. g. between the hours of 8 a.m. and 7 p.m. and average over the day.
2. Evaluate three-quarters of the optimum cycle calculated for the heaviest peak hour.
3. Select whichever is greater for the cycle time.

It is suggested as a reasonable procedure that the division of the available green time ($c_0 - L$) should be in proportion to the average y values for peak periods only, i.e.:

$$\frac{g_1}{g_2} = \frac{(y_1)_{\text{PEAK}}}{(y_2)_{\text{PEAK}}}$$

where $(y_1)_{\text{PEAK}}$ is the average y value during the peak periods for phase 1 and $(y_2)_{\text{PEAK}}$ that for phase 2.

Arterial Systems – Progression Models

An arterial is a road consisting of two or more intersections. A *signal system* is defined as having two or more individual signal installations which are linked together for coordination purposes. To obtain coordination all signals must operate with the same (common) cycle length, i.e., the signals must be synchronized. In rare instances, some intersections within the system may operate at double or one-half the cycle length of the system in which case there is half (or double-) synchronization. It is desirable that the arterial for which coordination is being provided have a green plus yellow interval equivalent to at least 50% of the cycle length. Two intersecting systems form an *open network* whenever they have only one intersection in common. A *closed network* is formed whenever there are three or more common intersections. A grid network of $m \times n$ arterials is a closed network with $m \times n$ common intersections. Closed networks are characterized by having *closed loops*.

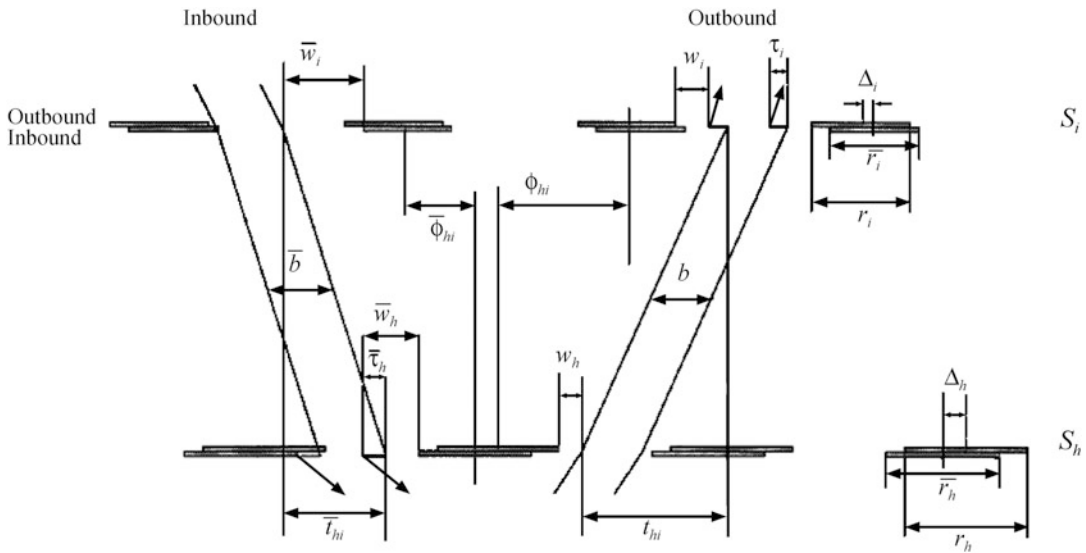
The predominant flow of traffic in urban street networks is along arterial streets. Optimal control of the traffic signals along these arteries is essential for the effective operation of the grid network.

Coordination of traffic lights along arterial streets provides numerous advantages:

1. A higher level of traffic service is provided in terms of higher overall speed and reduced number of stops.
2. Traffic flows more smoothly, often with an improvement in capacity due to reduced headways.
3. Vehicle speeds are more uniform because there is no incentive to travel at excessively high speeds to reach a signalized intersection within a green interval that is not in step. Also, the slow driver is encouraged to speed up in order to avoid having to stop for a red light.
4. There are fewer accidents because the platoons of vehicles arrive at each signal when it is green, thereby reducing the possibility of red signal violations and rear-end collisions.
5. Greater obedience to the signal commands is obtained from both motorists and pedestrians. The motorist tries to keep within the green interval and the pedestrian stays at the curb because the vehicles are more tightly spaced.
6. Through traffic tends to stay on the arterial streets instead of diverting onto parallel minor streets in search of alternative routes.

There are two basic approaches to computing arterial signal timings: (a) maximize the bandwidth of the progression, or (b) minimize overall delays and stops.

Arterial progression schemes are widely used for signal coordination. The conceptual basis for progression design is that traffic signals tend to group vehicles into a "platoon" with more uniform headways than would otherwise occur. The platooning effect is most prevalent on major streets with signalized intersections at frequent intervals. It is desirable, in these circumstances, to encourage platooning so that continuous movement (or progression) of vehicle platoons through successive traffic lights can be maintained. Signal timings, in this case, are designed to maximize the width of continuous green bands in both directions along the artery at the expected or recommended speed of travel. Although maximizing the bandwidth is only a surrogate performance measure, it provides significant benefits in increased



Optimization and Control of Urban Traffic Networks, Fig. 9 Time space diagram for MILP-1

travel speeds and smoothness of traffic flow and is the preferred approach in many countries. Such systems operate best when main-street flow is predominantly through traffic and when the number of vehicles turning onto the main street is small compared with the through volume. The second approach uses direct performance models to minimize delay, stops or other measures of disutility. Due to the complexity of these models and the high degree of non-linearity involved, they are generally not amenable to solution by mathematical programming methods and heuristic approaches have to be used. As a consequence, the resulting solutions are nonoptimal. In this section we concentrate on models of the first kind.

The first optimization model used for arterial bandwidth maximization was developed by Morgan and Little using a combinatorial optimization scheme (Morgan and Little 1964). A more advanced model, using mixed-integer linear programming, was later formulated by Little (1966). Advances in optimization techniques and computational capabilities have steadily increased the sophistication and versatility of the traffic signal optimization models. We introduce, first, single arterial bandwidth optimization models using both uniform and variable-width bandwidth progressions. These models are then extended to the optimization of arterial grid networks. Heuristic

decomposition approaches are developed to speed-up the computation and to make it suitable for on-line implementation as well as for combination with traffic assignment and routing models.

The models presented below use mixed-integer linear programming (MILP) for optimization. They are presented with progressively increasing complexity. MILP1 is the basic uniform-width bandwidth maximization model for a single arterial. MILP-2 introduces a multiband, multi-weight approach that generates variable width two-way progression bands along the arterial. MILP-3 extends the previous models to a grid network of intersecting arterials.

The Basic Bandwidth Maximization Problem

The geometric relations for the uniform bandwidth model are shown in Fig. 9. Consider an arterial having n signalized intersections. Let S_i denote the i th signal on the arterial and L_i denote the i th link (between signals i and $i + 1$) of the arterial, with $i = 1, \dots, n_j$. All time variables are defined in units of the cycle time. The following variables are defined:

C = cycle time (sec);

$b(\bar{b})$ = outbound (inbound) bandwidth;

$r_i(\bar{r}_i)$ = outbound (inbound) red time at S_i ;

$w_i(\bar{w}_i)$ = interference variables, time from right (left) side of red at S_i to left (right) side of outbound (inbound) green band;
 $t_{hi}(\bar{t}_{hi})$ = travel time from S_i to S_h in the outbound (inbound) direction;
 $d_{hi}(\bar{d}_{hi})$ = distance from S_i to S_h in the outbound (inbound) direction;
 $\phi_{hi}(\bar{\phi}_{hi})$ = internode offsets, time from the center of the outbound (inbound) red at S_h to the center of the outbound (inbound) red at S_i ;
 Δ_i = directional node phase shift, time from center of $\bar{r}_i$ to nearest center of r_i ;
 $\tau_i(\bar{\tau}_i)$ = queue clearance time for advancement of outbound (inbound) bandwidth at S_i to clear turning-in traffic before arrival of main-street platoon;
 $V_i(\bar{V}_i)$ = outbound (inbound) progression speed on link L_i (m/s).

In the case of uniform bandwidths the objective function has the following form:

$$\text{Maximize } b + k \cdot \bar{b} \quad (5)$$

where k is the target ratio of inbound to outbound band-width taken as the ratio of total inbound to total outbound volumes along the arterial.

The first set of constraints that we introduce are the *directional interference constraints*. These constraints make sure that the progression bands use only the available green time and do not infringe upon any of the red times. From Fig. 9 we see that:

$$w_i + b \leq 1 - r_i \quad (6a)$$

$$\bar{w}_i + \bar{b} \leq 1 - \bar{r}_i. \quad (6b)$$

Next, we calculate the *arterial-loop integer constraint*. This constraint is due to the fact that the signals of the arterial are synchronized, i.e., they operate in a common cycle time. If we start at the center of the outbound red at S_j and proceed along a loop consisting of the following points:

- Center of outbound red at S_i —
- Center of inbound red at S_i —
- Center of inbound red at S_h —
- Center of outbound red at S_h ,

we must end up at a point that is removed an integral number of cycle times from the point of departure. Summing algebraically the appropriate internode offsets and the directional node phase shift along the loop, we obtain

$$\phi_{hi} + \bar{\phi}_{hi} + \Delta_h - \Delta_i = v_{hi} \quad (7)$$

where v_{ij} is the corresponding loop integer variable.

From Fig. 9, we observe the following identities:

$$\phi_{hi} + \frac{r_i}{2} + w_i + \tau_i = \frac{r_h}{2} + w_h + t_{hi} \quad (8a)$$

and:

$$\bar{\phi}_{hi} + \frac{\bar{r}_i}{2} + \bar{w}_i + \bar{\tau}_i = \frac{\bar{r}_h}{2} + \bar{w}_h + \bar{t}_{hi}. \quad (8b)$$

Substituting (8b) into (8a) to eliminate ϕ and $\bar{\phi}$ gives:

$$\begin{aligned}
 & t_{hi} + \bar{t}_{hi} + \frac{r_i + \bar{r}_h}{2} + (w_h + \bar{w}_h) \\
 & - \frac{r_i + \bar{r}_i}{2} - (w_i + \bar{w}_i) - (\tau_i + \bar{\tau}_h) \\
 & + \Delta_h - \Delta_i = v_{hi}.
 \end{aligned} \quad (9)$$

So far we have assumed that S_i follows S_h in the outbound direction, but this restriction is not necessary. Let x represent any of the variables t , $\bar{t}$, v , ϕ , $\bar{\phi}$ then the following relationship is satisfied:

$$x_{hj} = x_{hi} + x_{ij}$$

whence, by setting $h = j$, we obtain $x_{hi} = -x_{ih}$. Thus Eqs. (8a) and (8b) hold for arbitrary S_h and S_i .

To simplify notation we number the signals sequentially from 1 to n in the outbound direction, and we define $x_i = x_{i,i+1}$. We can now rewrite Eq. (9) as follows:

$$\begin{aligned}
 & t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2} \\
 & - (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = v_i.
 \end{aligned} \quad (10)$$

The common signal cycle time C (seconds) and the link specific progression speed $V_i(\bar{V}_i)$ (m/sec)

are optimizable variables as well. This introduces considerable flexibility in the calculation of the best arterial progression. Each of these variables is constrained by upper and lower limits. In addition, changes in speed from one link to the next can also be limited. Let the limits be as follows:

C_1, C_2 = lower and upper bounds on cycle length;
 $(e_i, f_i), (\bar{e}_i, \bar{f}_i)$ = lower and upper bounds on
 outbound (inbound) speed $V_i(\bar{V}_i)$ (m/s);
 $(g_i, h_i), (\bar{g}_i, \bar{h}_i)$ = lower and upper bounds on
 change in outbound (inbound) speed $V_i(\bar{V}_i)$
 (m/sec).

To obtain constraints that are linear in the decision variables we define the inverse of the cycle time, i.e., the signal frequency $z = 1/C$ (cycles/second), such that:

$$1/C_2 \leq z \leq 1/C_1$$

The progression speeds and the changes in speeds are also expressed in terms of their reciprocals:

$$\frac{1}{f_i} \leq \frac{1}{V_i} \leq \frac{1}{e_i} \quad \text{and} \quad \frac{1}{\bar{f}_i} \leq \frac{1}{\bar{V}_i} \leq \frac{1}{\bar{e}_i} \quad \frac{1}{h_i} \leq \frac{1}{V_{i+1}} - \frac{1}{V_i} \leq \frac{1}{g_i}$$

$$\text{and} \quad \frac{1}{\bar{h}_i} \leq \frac{1}{\bar{V}_{i+1}} - \frac{1}{\bar{V}_i} \leq \frac{1}{\bar{g}_i}$$

Using the notation $d_i = d_{i,i+1}$, we obtain

$$t_i = \frac{d_i}{V_i} z \quad \text{and} \quad \bar{t}_i = \frac{\bar{d}_i}{\bar{V}_i} z$$

Substituting these expressions above and manipulating the variables, we obtain

$$\frac{d_i}{f_i} z \leq t_i \leq \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \leq \bar{t}_i \leq \frac{\bar{d}_i}{\bar{e}_i} z$$

$$\frac{d_i}{h_i} z \leq \frac{d_i}{d_{i+1}} t_{i+1} - t_i \leq \frac{d_i}{g_i} z \quad \text{and}$$

$$\frac{\bar{d}_i}{\bar{h}_i} z \leq \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \leq \frac{\bar{d}_i}{\bar{g}_i} z.$$

Thus, $t_i, \bar{t}_i$ and z are decision variables which, once known, determine the progression speeds. The complete MILP-1 is given below:

MILP-1. Find $b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i, v_i$ to
Maximize $b + k \cdot \bar{b}$ **subject to**

$$1/C_2 \leq z \leq 1/C_1$$

$$w_i + b \leq 1 - r_i \quad \text{and} \quad \bar{w}_i + \bar{b} \leq 1 - \bar{r}_i$$

$$i = 1, \dots, n$$

$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2}$$

$$- (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = v_i$$

$$i = 1, \dots, n-1$$

$$\frac{d_i}{f_i} z \leq t_i \leq \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \leq \bar{t}_i \leq \frac{\bar{d}_i}{\bar{e}_i} z$$

$$i = 1, \dots, n-1$$

$$\frac{d_i}{h_i} z \leq \frac{d_i}{d_{i+1}} t_{i+1} - t_i \leq \frac{d_i}{g_i} z \quad \text{and}$$

$$\frac{\bar{d}_i}{\bar{h}_i} z \leq \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \leq \frac{\bar{d}_i}{\bar{g}_i} z \quad i = 1, \dots, n-1$$

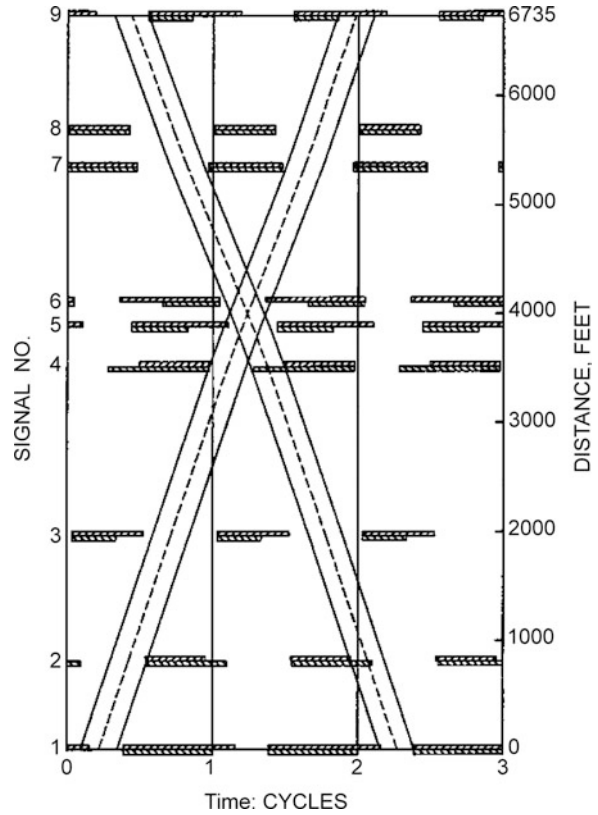
$$b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i \geq 0 \quad \text{and} \quad v_i \text{ integer.}$$

In addition to bandwidths and interference variables, this program also calculates optimal cycle length ($1/z$) and progression speeds. To calculate the offsets, green splits need to be given or calculated based on flow/capacity considerations. Phase sequencing can also be optimized. MILP-1 is the model used by the MAXBAND program developed by Little et al. (1982). An example of a uniform bandwidth solution obtained by MILP-1 is shown in Fig. 10. The scheme contains signals with multi-phase sequences.

The Variable Bandwidth Problem

In this section we describe a more versatile multi-band/multi-weight approach for the bandwidth maximization problem. A different bandwidth is defined for each directional road section of the arterial and is individually weighted with respect to its contribution to the overall objective function. While the band is continuous, its width can vary and adapt to the prevailing traffic flow on each link. In this way the formulation is sensitive to varying traffic conditions and can tailor the progression scheme to the different possible traffic flow patterns. The user can still choose a uniform bandwidth progression if desired, but this is only one of the many possible options. Referring to the geometry in Fig. 11, we define the following variables:

Optimization and Control of Urban Traffic Networks, Fig. 10 Time space diagram for a uniform bandwidth progression scheme generated by MILP-1



b_i ($\bar{b}_i$) = outbound (inbound) bandwidth of link i ;
 there is now a specific band for each link L_i ;
 w_i ($\bar{w}_i$) = the time from right (left) side of red at S_i
 to the centerline of the outbound (inbound)
 green band; the reference point at each signal
 has been moved from the edges to the center-
 line of the band.

The following constraints apply in the out-
 bound directions at signal S_i :

$$w_i + \frac{b_i}{2} \leq 1 - r_i \quad \text{and} \quad w_i + \frac{b_i}{2} \geq 0.$$

The pair of constraints can be combined as
 follows:

$$\frac{b_i}{2} \leq w_i \leq (1 - r_i) - \frac{b_i}{2}.$$

A similar relation must be observed at S_{i+1} ,
 since band b_i must be constrained at both ends:

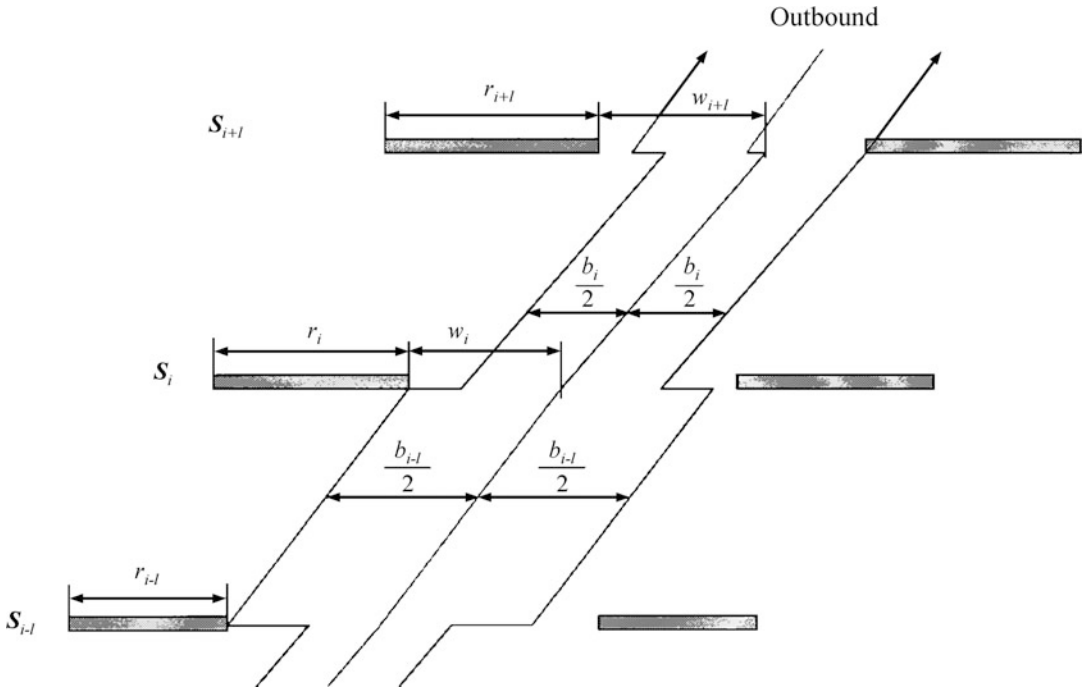
$$\frac{b_i}{2} \leq w_{i+1} \leq (1 - r_{i+1}) - \frac{b_i}{2}.$$

Corresponding relationships exist in the
 inbound direction (marked by a bar on all
 variables):

$$\begin{aligned} \frac{\bar{b}_i}{2} &\leq \bar{w}_i \leq (1 - \bar{r}_i) - \frac{\bar{b}_i}{2} \\ \frac{\bar{b}_i}{2} &\leq \bar{w}_{i+1} \leq (1 - \bar{r}_{i+1}) - \frac{\bar{b}_i}{2}. \end{aligned}$$

These constraints are referred to as the *directional interference constraints*. The time reference points are redefined to the *centerline* of the bands (or, the *progression line*), rather than the edges. The *loop integer constraint* given by Eq. (10) remains unchanged and so do the travel time and speed-change constraints.

The most important change occurs in the objective function. Since the bands are link-specific, they can be weighted disaggregately to reflect user-defined traffic performance objectives



Optimization and Control of Urban Traffic Networks, Fig. 11 Geometric relations for MILP-2

which can be individually weighted for each link of the arterial. The objective function now has the following form:

$$\text{Maximize } \frac{1}{n-1} \sum_{i=1}^n a_i \cdot b_i + \bar{a}_i \cdot \bar{b}_i \quad (11)$$

where a_i and $\bar{a}_i$ are the link specific weighting coefficients for the outbound and inbound directions respectively. There is a multitude of possible options for choosing the weighting coefficients in Eq. (11). Some of the more common expressions are as follows:

$$a_i = \left(\frac{q_i}{s_i} \right)^p \quad \text{and} \quad \bar{a}_i = \left(\frac{\bar{q}_i}{\bar{s}_i} \right)^p$$

where

$q_i(\bar{q}_i)$ = outbound (inbound) directional flow rate on link L_i ; either the total or the through volume can be used;

$s_i(\bar{s}_i)$ = saturation flow rate outbound (inbound) directional volume on link V_{ij} ; either the total flow rate or the through flow rate can be used;

p = an integer exponent; the values 0, 1, 2 and 4 were used.

To obtain an objective function value that is consistent with those used previously, we normalize the weighting coefficients to obtain

$$\sum_{i=1}^{n-1} a_i = n-1 \quad \text{and} \quad \sum_{i=1}^{n-1} \bar{a}_i = n-1$$

The multi-band/multi-weight optimization program is summarized in MILP-2.

MILP-2. Find $b_i, \bar{b}_i, z, w_i, \bar{w}_i, t_i, \bar{t}_i, v_i$ to

$$\text{Maximize } \frac{1}{n-1} \sum_{i=1}^n a_i \cdot b_i + \bar{a}_i \cdot \bar{b}_i \text{ subject to}$$

$$1/C_2 \leq z \leq 1/C_1$$

$$\frac{b_i}{2} \leq w_i \leq (1-r_i) - \frac{b_i}{2}, \quad \frac{b_i}{2} \leq w_{i+1}$$

$$\leq (1-r_{i+1}) - \frac{b_i}{2},$$

$$\frac{\bar{b}_i}{2} \leq \bar{w}_i \leq (1-\bar{r}_i) - \frac{\bar{b}_i}{2} \quad \text{and}$$

$$\frac{\bar{b}_i}{2} \leq \bar{w}_{i+1} \leq (1-\bar{r}_{i+1}) - \frac{\bar{b}_i}{2} \quad i = 1, \dots, n-1$$

$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2}$$

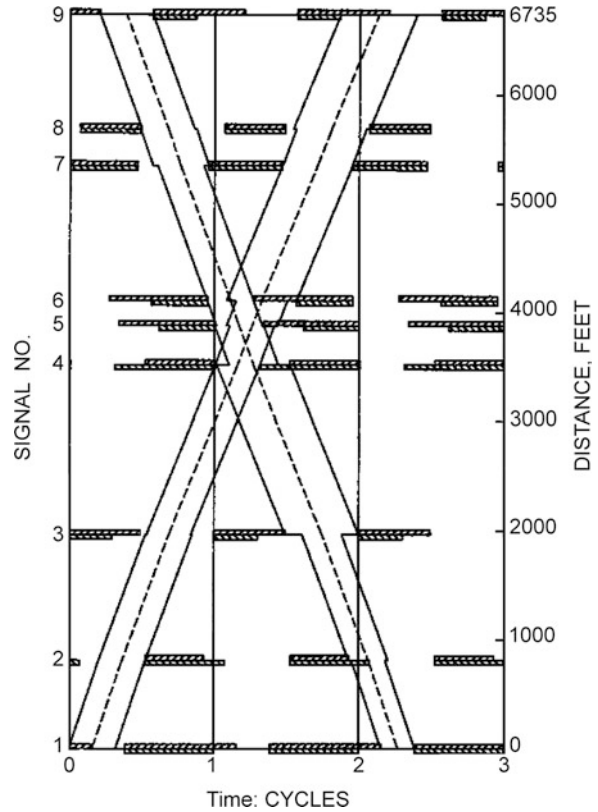
$$-(w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = v_i$$

$$i = 1, \dots, n-1$$

$$\begin{aligned}
\frac{d_i}{f_i}z \leq t_i \leq \frac{d_i}{e_i}z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i}z \leq \bar{t}_i \leq \frac{\bar{d}_i}{\bar{e}_i}z \\
i = 1, \dots, n-1 \\
\frac{d_i}{h_i}z \leq \frac{d_i}{d_{i+1}}t_{i+1} - t_i \leq \frac{d_i}{g_i}z \quad \text{and} \\
\frac{\bar{d}_i}{\bar{h}_i}z \leq \frac{\bar{d}_i}{\bar{d}_{i+1}}\bar{t}_{i+1} - \bar{t}_i \leq \frac{\bar{d}_i}{\bar{g}_i}z \quad i = 1, \dots, n-1 \\
b_i, \bar{b}_i, z, w_i, \bar{w}_i, t_i, \bar{t}_i \geq 0 \quad \text{and} \quad v_i \text{ integer.}
\end{aligned}$$

MILP-2 involves $(18n - 20)$ constraints and $(6n - 3)$ continuous variables, and $n - 1$ unrestricted integer variables, not counting slack variables. Green or red splits are determined the same as in MILP-1. The increased decision capabilities of MILP-2 require a corresponding increase in the size of the mathematical program. The formulation given in MILP-2 corresponds to the MULTIBAND optimization program developed by Gartner et al. (1991). The variable bandwidth solution for the same arterial as for MILP-1 (Fig. 9) is shown in Fig. 12.

Optimization and Control of Urban Traffic Networks, Fig. 12 Time space diagram for a variable bandwidth progression scheme calculated by MILP-2



The Network Problem

The MILP-2 problem can be extended for controlling traffic flow in a network of intersecting arterials. This represents a significantly more challenging problem: progressions must be provided on all the arterials of the network, simultaneously. Consider a network of m arterials with each arterial having n_j signalized intersections. Let S_{ij} denote the i th signal on the j th arterial of the network and L_{ij} denote the i th link (between signals i and $i + 1$) of the j th arterial, with $j = 1, \dots, m$ and $i = 1, \dots, n_j$. The variables presented in MILP-1 and MILP-2 are redefined as follows:

- $b_{ij}(\bar{b}_{ij})$ = outbound (inbound) bandwidth of link i on arterial j ;
- $w_{ij}(\bar{w}_{ij})$ = interference variables, the time from right (left) side of red at S_{ij} to the centerline of the outbound (inbound) green band;
- $r_{ij}(\bar{r}_{ij})$ = outbound (inbound) red time at S_{ij} ;

$t_{ij}(\bar{t}_{ij})$ = travel time on link i of arterial j in the outbound (inbound) direction;

$d_{ij}(\bar{d}_{ij})$ = length of link i of arterial j in the outbound (inbound) direction;

$\phi_{(i,j),(il)}(\bar{\phi}_{(i,j),(kj)})$ = internode offsets, time from the center of the outbound (inbound) red at S_{ij} to the center of the outbound (inbound) red at S_{kj} ;

Δ_{ij} = directional node phase shift, time from center of $\bar{r}_{ij}$ to nearest center of r_{ij} ;

$\tau_{ij}(\bar{\tau}_{ij})$ = queue clearance time for advancement of outbound (inbound) bandwidth at S_{ij} to clear turning-in traffic before arrival of main-street platoon;

$V_{ij}(\bar{V}_{ij})$ = outbound (inbound) progression speed on link L_{ij} (m/s).

$e_{ij}, f_{ij}, (\bar{e}_{ij}, \bar{f}_{ij})$ = lower and upper bounds on outbound (inbound) speed $V_{ij}(\bar{V}_{ij})$ (m/s);

$g_{ij}, h_{ij}, (\bar{g}_{ij}, \bar{h}_{ij})$ = lower and upper bounds on change in outbound (inbound) speed $V_{ij}(\bar{V}_{ij})$ (m/s).

In addition to the above variables, the *intranode offset* is defined as:

$\omega_{(ij),(kl)}(\bar{\omega}_{(ij),(kl)})$ = intranode offsets, the time from the center of the inbound (outbound) red at i th node of the j th arterial (S_{ij}), to the center of the inbound (outbound) red in the crossing direction at the same node which is also the k th node of the l th arterial (S_{kl}).

The directional interference constraints, the arterial loop integer constraint and the travel time and speedchange constraints remain unchanged and are applied, in turn, for each arterial. A new set of constraints, though, is required for the synchronization of all signals. Similar to the arterial loop-integer constraints Eq. (7), for any closed loops of the network consisting of more than 2 links, the summation of *internode* and *intranode offsets* around the loop of intersecting arterials must be an integer multiple of the cycle time. This results in the formulation of the *network loop-integer constraints*. The number of network loop constraints and the choice of a fundamental set of loops is given by Gartner (1972).

The network optimization problem is summarized in MILP-3:

MILP-3. Find $b_{ij}, \bar{b}_{ij}, z, w_{ij}, \bar{w}_{ij}, t_{ij}, \bar{t}_{ij}, v_{ij}, \mu_i$ **to**

Maximize $\sum_{j=1}^m \frac{1}{n-1} \sum_{i=1}^n a_{ij} \cdot b_{ij} + \bar{a}_{ij} \cdot \bar{b}_{ij}$ **subject to**

$$1/C_2 \leq z \leq 1/C_1$$

$$\frac{b_{ij}}{2} \leq w_{ij} \leq (1 - r_{ij}) - \frac{b_{ij}}{2},$$

$$\frac{b_{ij}}{2} \leq w_{i+1,j} \leq (1 - r_{i+1,j}) - \frac{b_{ij}}{2},$$

$$\frac{\bar{b}_{ij}}{2} \leq \bar{w}_{ij} \leq (1 - \bar{r}_{ij}) - \frac{\bar{b}_{ij}}{2} \quad \text{and}$$

$$\frac{\bar{b}_{i,j}}{2} \leq \bar{w}_{i+1,j} \leq (1 - \bar{r}_{i+1,j}) - \frac{\bar{b}_{ij}}{2}$$

$$i = 1, \dots, n_j - 1; j = 1, \dots, m$$

$$t_{ij} + \bar{t}_{ij} + \frac{r_{ij} + \bar{r}_{ij}}{2} + (w_{ij} + \bar{w}_{ij}) - \frac{r_{i+1,j} + \bar{r}_{i+1,j}}{2} - (w_{i+1,j} + \bar{w}_{i+1,j})$$

$$- (\tau_{i+1,j} + \bar{\tau}_{i,j}) + \Delta_{i,j} - \Delta_{i+1,j} = v_{i,j}$$

$$i = 1, \dots, n_j - 1; \quad j = 1, \dots, m$$

$$\phi_{(ia),(i+1,a)} + \omega_{(i+1,a),(jb)} + \phi_{(jb),(j+1,b)} + \omega_{(j+1,b),(kc)} + \phi_{(kc),(k+1,c)}$$

$$+ \omega_{(k+1,c),(ld)} + \phi_{(ld),(l+1,d)} + \omega_{(ld),(ia)} = \mu_N$$

$$i = 1, \dots, n_a - 1; \quad j = 1, \dots, n_b - 1;$$

$$k = 1, \dots, n_c - 1; \quad l = 1, \dots, n_d - 1;$$

a, b, c, d arterials forming a fundamental network loop;

$$\frac{d_{ij}}{f_{ij}} z \leq t_{ij} \leq \frac{d_{ij}}{e_{ij}} z \quad \text{and} \quad \frac{\bar{d}_{ij}}{\bar{f}_{ij}} z \leq \bar{t}_{ij} \leq \frac{\bar{d}_{ij}}{\bar{e}_{ij}} z$$

$$i = 1, \dots, n_j - 1; j = 1, \dots, m$$

$$\frac{d_{ij}}{h_{ij}} z \leq \frac{d_{ij}}{d_{i+1,j}} t_{i+1,j} - t_{ij} \leq \frac{d_{ij}}{g_{ij}} z \quad \text{and}$$

$$\frac{\bar{d}_{ij}}{h_{ij}} z \leq \frac{\bar{d}_{ij}}{d_{i+1,j}} \bar{t}_{i+1,j} - \bar{t}_{ij} \leq \frac{\bar{d}_{ij}}{g_{ij}} z$$

$$i = 1, \dots, n_j - 1; j = 1, \dots, m$$

$$b_{ij}, \bar{b}_{ij}, z, w_{ij}, \bar{w}_{ij}, t_{ij}, \bar{t}_{ij} \geq 0 \quad \text{and}$$

$$v_{ij}, \mu_i \text{ integer.}$$

For an $m \times n$ closed grid network ($m \times n$ intersections and $m + n$ arterials), MILP-3 may involve up to $(23mn - 14m - 14n + 1)$ constraints and $(12mn - 6m - 6n)$ continuous variables, and $(3mn - 2m - 2n + 1)$ unrestricted integer variables, not counting slack variables.

The principal difficulty in solving mixed-integer problems, such as the ones described above, is the number of integer variables and their range, i.e., the size of the integer feasible set. Virtually all MILP codes use branch-and-bound strategies for calculating the optimal values. Solving the primary multi-band problem by general purpose branch-and-bound is still a rather formidable task. For this reason, heuristic methods which quickly lead to a good solution have been developed (Gartner and Stamatiadis 2002).

Delay-Based Models – Cyclic Flow Profiles

TRANSYT

TRANSYT, the traffic network study tool, is a computer model to optimize traffic signal timings and perform traffic signal simulation. TRANSYT has two main elements – the traffic model which is used to calculate the performance index of the traffic network for a given set of signal timings and an optimizing process that makes changes to the settings and determines whether they improve the performance index or not.

Simulation Model

The traffic simulation model used in TRANSYT is a macroscopic traffic simulation model. In a macroscopic model platoons of vehicles are considered rather than individual vehicles. TRANSYT simulates traffic flow macroscopically, but in a step-wise manner. The cycle length is divided into small, equal time increments, called steps. All TRANSYT calculations are made on the basis of flow rates (such as vehicles per hour). The average flow pattern past a point in the road network is represented by a histogram also known as Cyclic Flow Profile. There are three traffic flow patterns considered while simulating behavior of a signalized intersection: the “IN,” “GO,” and “OUT” patterns (see figure below) (Robertson 1969).

The IN-Pattern or Arrival Flow Pattern

The IN-Pattern is the pattern of traffic arriving at the stop line. The arrival flow pattern is expressed mathematically as follows:

$$IN_{it} = \sum_j^n F_{ij}(P_{ij} \cdot OUT_{jt'})$$

where,

IN_{it} = the IN-pattern on link i for time step t

F_{ij} = smoothing process related to platoon dispersion for flow to link i from link j

P_{ij} = the proportion of the feeding link OUT-pattern that feeds the subject link

$OUT_{jt'}$ = the OUT-pattern of link j for step t'

t' = the time step t minus the travel time for flow to link i from link j

n = the number of links (j) that feed link i

The IN-pattern is computed for each step, t , in the cycle, thus forming a pattern, or flow profile (Fig. 13).

The GO-Pattern or Saturation Flow Pattern

The GO-pattern is the flow rate at each step that would leave the stop line when the traffic signal turns green if there were enough traffic to saturate the green. There is a certain Startup Lost Time (SLT) involved due to reaction of driver etc.

The OUT-Pattern or Departure Flow Pattern

The OUT-pattern is the profile of traffic actually leaving the stop line. It is usually equal to the GO-pattern as long as there is a queue. After the queue dissipates, it is equal to the IN-pattern for the remainder of the effective green.

The queue (or the number of vehicles held at the stop line during any time interval, t) is first determined to determine the OUT-pattern

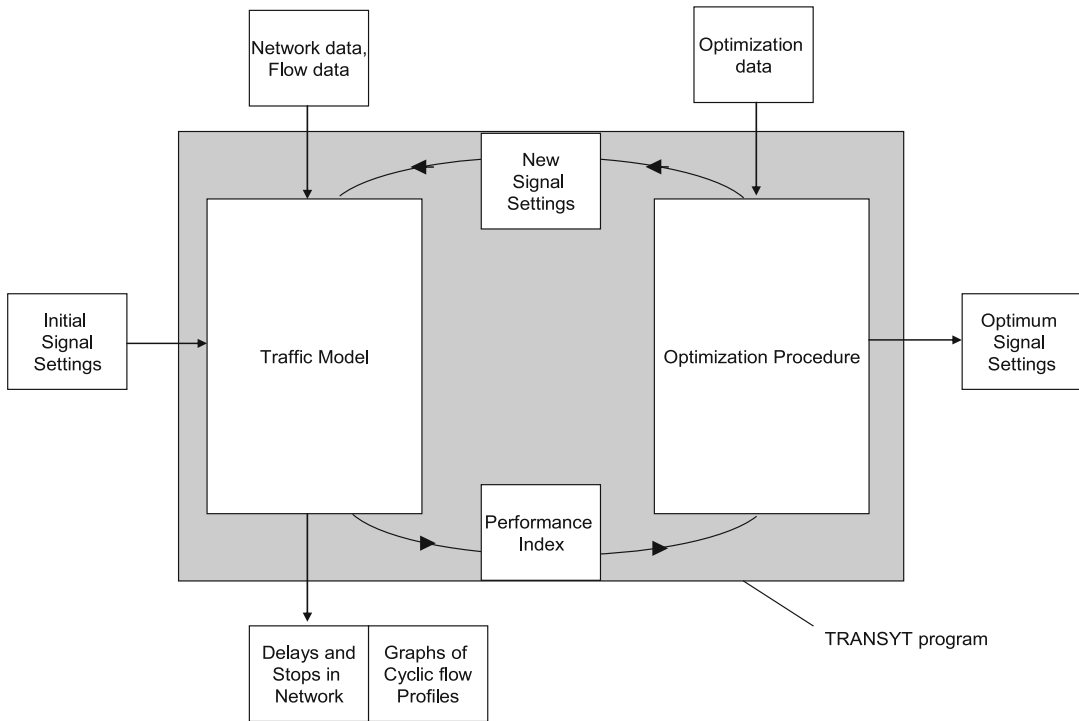
$$m_t = \max\{(m_{t-1} + q_t - s_t), 0\}$$

where,

m_t = number of vehicles in the queue in time interval t on a given link (and similarly for m_{t-1})

q_t = number of vehicles arriving in interval t , given by the IN-pattern

s_t = number of vehicles allowed to leave in interval t , given by the GO-pattern



Optimization and Control of Urban Traffic Networks, Fig. 13 Structure of TRANSYT program

The OUT-pattern is given for link i during time interval t by the expression:

$$\text{OUT}_{it} = m_{i,t-1} + q_{it} - m_{it}.$$

The traffic model further utilizes a platoon dispersion algorithm that simulates the normal dispersion (i.e., the “spreading out”) of platoons as they travel downstream. The start-and-stop operation of signals tends to create platoons of vehicles that travel along a link. TRANSYT models the dispersion of these platoons as they progress along a link.

On internal links, for each time interval (step), t , the downstream arrival flow is determined by the following recurrence equation:

$$v'_{(t+\beta T)} = F \cdot v_t + \left[(1 - F) \cdot v'_{(t+\beta T-1)} \right]$$

where,

$v'_{(t+\beta T)}$ = predicted flow rate in time interval $t + \beta T$ of the predicted platoon

v_t = flow rate of the initial platoon during step t

β = factor to take into account vehicles traveling faster than speed limit

T = the cruise travel time on the link, in steps

F = a smoothing factor, where:

$$F = (1 + \alpha \beta T)^{-1},$$

where α = an empirically derived constant, called the platoon dispersion factor (PDF) takes into account site-specific factors such as grade, curvature, parking, opposing flow interference, and other sources of impedance.

Optimization Model

TRANSYT uses a Performance Index (PI) (also the objective function) allowing the user to define their preferences regarding performance of the traffic network. TRANSYT develops a signal timing plan that produces an optimal value of the PI. TRANSYT-7F (the US-adopted version) offers numerous PI's to choose from, which can

reflect anything that the user desires including delay, progression, stops, fuel consumption, queuing, and throughput. These PI options are listed below:

$$PI = \begin{bmatrix} \text{DI only} \\ \text{PROS only} \\ \text{PROS\&DI} \\ \text{PROS/DI} \\ \text{DI} \sum \frac{\text{Average back of queue on link } i}{\text{Queuing capacity on link } i} \\ \text{Throughput only} \\ \text{Throughput\&DI} \\ \text{Throughput/DI} \end{bmatrix}$$

In general, TRANSYT-7F always attempts to maximize the PI, unless the disutility index (DI) has been selected as the PI. Since the DI can be a combination of delay, stops, queuing, and fuel consumption, this value must be minimized. The other objective functions that involve progression or throughput must be maximized.

The DI is a combination of vehicle delay, stops, and fuel consumption. Weighting factors are available in order to place more emphasis on any of these three DI components if desired. Delay-only optimization results in excessive stops and fuel consumption. Excess fuel consumption (minimization) is considered to be a good compromise between bandwidth and delay-based optimization.

The “standard” delay and stops DI optimization objective function is defined as follows:

$$DI = \sum_{i=1}^n \{ (w_{di} d_i + K w_{si} S_i) + QP \}$$

where,

DI = disutility index, analogous to the original TRANSYT performance index

d_i = delay on link i (of n links) and on an optional userspecified upstream input link $i - 1$

K = user-coded stop penalty factor to express the importance of stops relative to delay

S_i = stops on link i per second

w_{xi} = link-specific weighting factors for delay (d) and stops (s) on link i

U_i = binary variable that is ‘1’ if link-to-link weighting has been established, zero otherwise
QP = queuing penalty

A queuing penalty is used to minimize the possibility of spillback.

$$QP = QB_i W_q (q_i - qc_i)^2$$

where,

Q = a binary variable, ‘1’ if the queue penalty is included in the DI ‘0’ otherwise

B_i = a binary variable, ‘1’ if the maximum back of queue (q_i) exceeds the queuing capacity ‘0’ otherwise

W_q = a network-wide penalty applied to the excess queue spillback

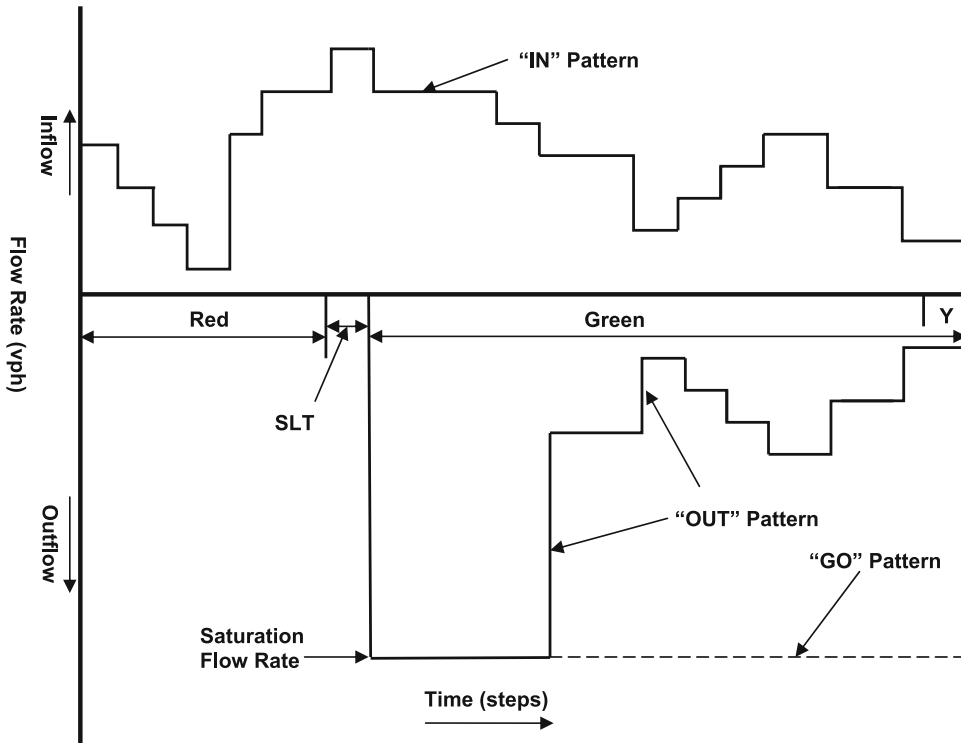
q_i = computed maximum back of queue on link i
 qc_i = queuing capacity for link i

Progression opportunities (PROS) represent the ability of vehicles to progress through multiple intersections without stopping. Points are scored every time a vehicle is able to traverse additional intersections. Therefore it is possible to achieve PROS without having a wide bandwidth stretching from the beginning of an arterial street to the end (Fig. 14).

It is also possible to select a combination of PROS and DI in order to achieve a compromise between the two objectives. For example, if the PI is defined as PROS/DI, the program attempts to simultaneously maximize the numerator (PROS) and minimize the denominator (DI). For optimizing bandwidth and progression speeds, PROS/DI is recommended. PROS-only is rarely acceptable, because although it often produces a good bandwidth, the minor movements are allowed to fail.

Demand-Responsive Models

A number of demand-responsive, or adaptive signal control strategies have been developed. Two such strategies are described herein: SCOOT and OPAC.



Optimization and Control of Urban Traffic Networks, Fig. 14 TRANSYT flow patterns

SCOOT

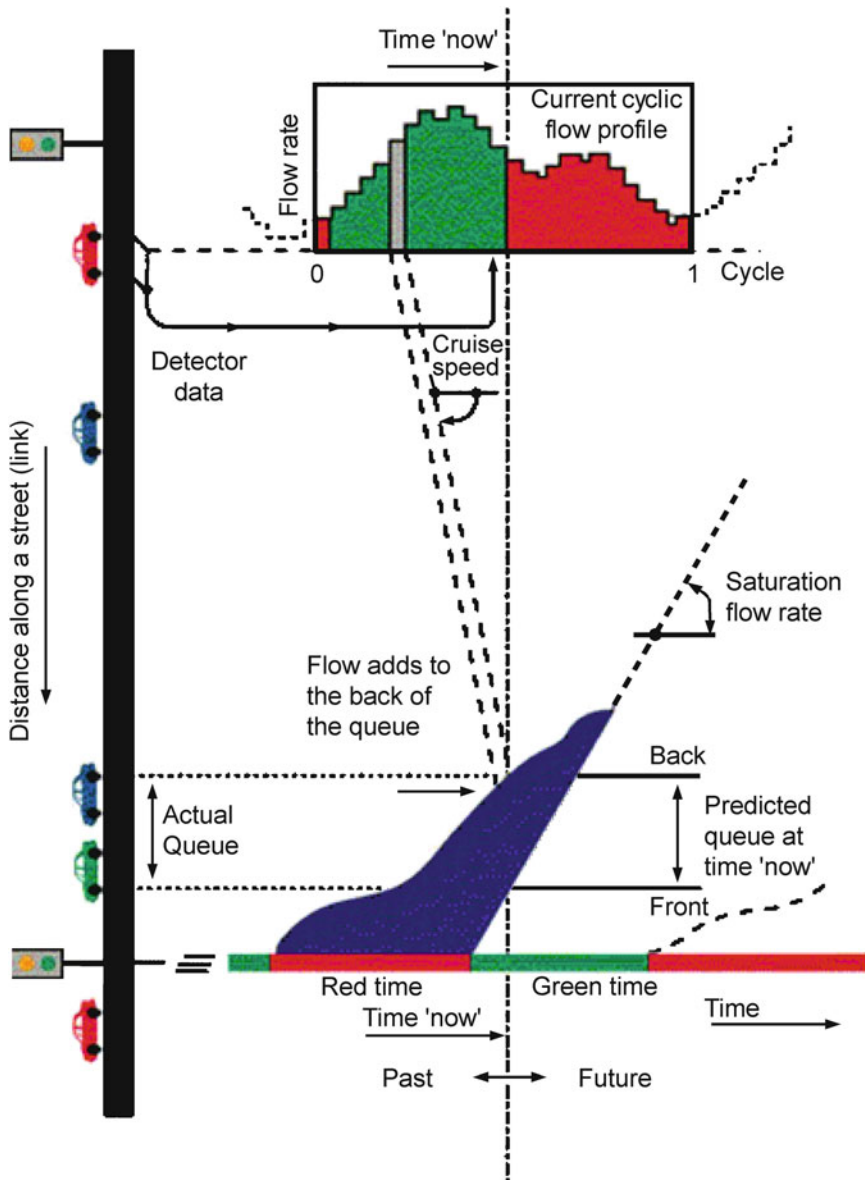
Adaptive traffic signal systems adjust signal timings based on measured flow. Reacting to these flow variations results in reduced delay, shorter queues and decreased travel time. SCOOT (Split Cycle Offset Optimization Technique) is one of the earliest real time adaptive traffic control systems (Hunt et al. 1981). It coordinates the operation of all the traffic signals in an area to give good progression to vehicles through the network.

The operation of the SCOOT model is illustrated in Fig. 15. SCOOT obtains information on traffic flows from detectors. As an adaptive system, SCOOT depends on good traffic data so that it can respond to changes in flow. Detectors are normally required on every link. Their location is important and they are usually positioned at the upstream end of the approach link.

When vehicles pass the detector, SCOOT receives the information and converts the data into its internal units and uses them to construct "Cyclic flow profiles" for each link. The sample profile

shown in the diagram is color coded green and red according to the state of the traffic signals when the vehicles will arrive at the stop line at normal cruise speed. Vehicles are modeled down the link at cruise speed and join the back of the queue (if present). During the green, vehicles discharge from the stop line at the validated saturation flow rate.

The data from the model is then used by SCOOT in three optimizers which are continuously adapting three key traffic control parameters – the amount of green for each approach (Split), the time between adjacent signals (Offset) and the time allowed for all approaches to a signaled intersection (Cycle time). These three optimizers are used to continuously adapt these parameters for all intersections in the SCOOT controlled area, minimizing wasted green time at intersections and reducing stops and delays by synchronizing adjacent sets of signals. This means that signal timings evolve as the traffic situation changes without any of the harmful disruption caused by changing fixed time plans on more traditional urban traffic control systems.



Optimization and Control of Urban Traffic Networks, Fig. 15 Schematic diagram showing operation of SCOOT model

OPAC – Optimization Policies for Adaptive Control

This section presents an adaptive control strategy for cocoordinating and synchronizing signals in a network using the virtual-fixed-cycle concept and is labeled *VFCOPAC*. The strategy is based on the single intersection dynamic-programming-based *OPAC* (optimization policies for adaptive control)

adaptive controller developed by Gartner (1983). *VFC-OPAC* consists of a distributed control strategy featuring a dynamic optimization algorithm that calculates signal timings to minimize a performance function of total intersection delays and stops. The algorithm uses a combination of measured and modeled demand to determine, in a distributed manner, phase durations at each signal

that are constrained by minimum and maximum green times and, when running in a coordinated mode, by coordination and synchronization parameters that can be updated based on real-time data.

OPAC was developed from the outset as a distributed strategy featuring a dynamic optimization algorithm for traffic signal control without requiring a rigid, fixed cycle time. Signal timings are calculated to directly minimize performance measures, such as vehicle delays and stops, and are only constrained by minimum and maximum phase lengths and, if running in a coordinated mode, by a virtual cycle length and by a virtual offset. Development of the strategy has progressed through several versions, each one serving as a base for a subsequent version. The principal features of each version are outlined below.

OPAC I: Dynamic Programming

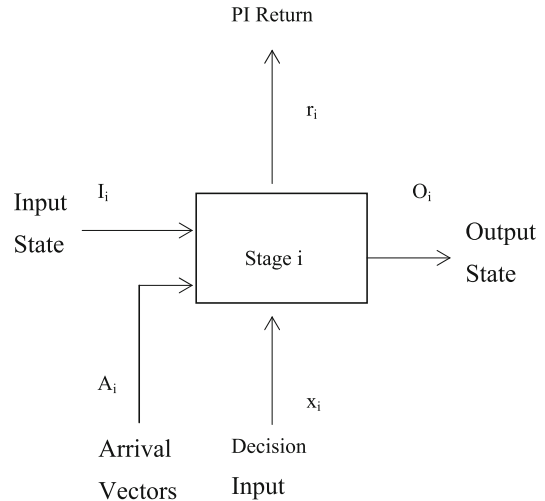
The first version, designated *OPAC I*, was designed to serve as a basis for subsequent *opac* strategy development. *OPAC I* utilizes a dynamic programming (DP) model for the determination of the traffic control parameters. Since DP is a global optimization strategy for multistage decision processes, it provides a standard against which all other strategies can be compared (Bellman and Dreyfus 1962; Grafton and Newell 1967).

The optimization process is decomposed into N stages, where each stage corresponds to a discrete time interval in which the arrivals are measured, (2 to 5 s intervals). The total number of stages N corresponds to the *horizon length* (HL) of the input predictions. For exploratory purposes the horizon length was assumed to be several minutes long, e. g., 5, 10, or 15 min. A typical stage i is illustrated in Fig. 16.

At stage i we have an input state vector I_i , an arrival vector A_i , output state vector O_i , input variable x_i , economic return (cost) output r_i , and a set of transformations:

$$\begin{aligned} O_i &= T_i(I_i, A_i, x_i) \\ r_i &= R_i(I_i, A_i, x_i). \end{aligned}$$

The state of the intersection is characterized by the state of the signal (green or red) and by the



Optimization and Control of Urban Traffic Networks,
Fig. 16 Dynamic Programming stage in *OPAC I*

queue-length on each of the approaches. The input decision variable indicates whether the signal is to be switched at this stage or to remain in its present state. The return cost output is the intersection's index of performance (e. g., total delay time and/or number of stops), which has to be minimized. The functional relationships between the input and output variables are based on the queuing-discharge process occurring at the intersection, i.e., the vehicle inflow and outflow as a function of the signal settings.

We define the following variables (all corresponding to stage i):

- a = approach designation, by direction, $a = N, S, E, W$
- A^a = number of arrivals during the stage (for a four-leg intersection)
- D^a = number of departures (discharges) during the stage
- Q^a = queue-length on approach at beginning of stage
- S^a = status of signal at start of stage (green/red)
- x = stage input decision variable (change/no-change)

The input state vector (transposed) is: $I^t = [I^N, I^S, I^E, I^W]$, where for each approach the state I

$$I^a = [S^a, Q^a]^t.$$

The output state vector contains the same elements as the input state vector and equals to the input state of the succeeding stage, i.e., $O_i = I_{i+1}$. The signal status for each approach a , is a binary variable:

$$S^a = \{0 \text{ for Green; } 1 \text{ for Red}\}$$

The input decision variable is also binary

$$x = \begin{cases} 0 & \text{no change in signal status} \\ 1 & \text{change current signal status} \end{cases}$$

The transformation of input to output, at stage i , is as follows:

$$\begin{aligned} S_{i+1}^a &= (S_i^a + x_i)_{\text{mod } 2} \\ Q_{i+1}^a &= Q_i^a + A_i^a - D_i^a \end{aligned}$$

The mod 2 operator ensures that S_{i+1}^a will always be 0 or 1. The arrivals at each stage i , A_i^a are an observed input (e. g., from detectors); the departures are a function of the state and decision variables:

$$D = \begin{cases} 0 & \text{if } S = 1 \\ \min(Q + A, d_{\max}) & \text{if } S = 0 \end{cases}$$

where $d_{\max}$ is the saturation discharge rate (in veh/int). The DP algorithm goes backward in time, i.e., starting from the last interval and back-tracking to the first, at which time an optimal switching policy for the entire horizon is determined. The switching policy consists of the sequence of phase switch-ons and switch-offs throughout the horizon. The recursive optimization functional is:

$$f_i^*(I_i) = \min_{x_i} \{R_i(I_i, A_i, x_i) + f_{i+1}^*(I_i, A_i, x_i)\}.$$

The Performance Index (return) at stage i is the queuing delay and number of stops N_s incurred at this stage:

$$\begin{aligned} r_i &= R_i(I_i, A_i, x_i) \\ &= \sum_a (Q_i^a + A_i^a - D_i^a) + \sum_a N_s^a. \end{aligned}$$

When the optimization is terminated at stage $i = 1$ we have,

$$\begin{aligned} f_1^*(I_1) &= \min_{x_i} \left\{ \sum_{i=1}^N R_i(I_i, A_i, x_i) \right\} \\ &= \sum_{i=1}^N R_i(I_i, A_i, x_i^*) \end{aligned} \quad (12)$$

which is the minimized Performance Index over the horizon period for a given initial input state I_1 . Since the initial conditions at stage 1 are specified (i.e., the queue lengths on all approaches are given as well as the initial signal status), we can retrace the optimal policy by taking a forward pass through the arrays of $X_i^*(I_1)$. The policy consists of the optimal sequence of switching decisions $\{x_i^*, i = 1, \dots, N\}$ at all stages of the optimization process.

While this procedure assures globally optimal controls for the given horizon length, it requires complete information of arrivals over the entire control period. It cannot be used for real-time implementation due to both the (excessive) amount of processing involved and due to the lack of a practical method to gather real-time information for such a length of time. Much of the output generated by the procedure is never implemented because optimized policies are calculated for all possible combinations of initial conditions at each stage of the control period. In practice, only one 'optimum policy' is implemented. Nonetheless, *OPAC-I* serves an important function as a standard for the evaluation of the relative effectiveness of other, more practical strategies.

OPAC II: Sequential Optimization

OPAC II breaks up the horizon into sequential optimization stages to speed up the optimization process. The model is a reformulation of the *OPAC I* algorithm. The purpose is to create a building block for a distributed online strategy. *OPAC II* has the following features:

- The control period is divided into successive (back-to back) horizon lengths of T seconds each (T may encompass one or more cycle lengths).

- Each horizon is divided into an integral number of intervals 't' seconds long; typically, $t = 2-5$ s.
- During each horizon there must be a sufficient number of phase changes to guarantee that no optimal solution is missed. The phase change (switching) times are measured from the start of the horizon in time units of t.
- For any given switching sequence in a horizon, the performance function for each approach computes the total delay and/or stops.

The optimization problem in OPAC II can be stated as follows: For each horizon length, given the initial queues on each approach and the arrivals for each interval of the horizon, determine the sequence of switching times, in terms of intervals, which yield the least delay and/or stops to vehicles over the entire horizon.

The procedure used for solving the problem consists of an intelligent search over the set of all possible combinations of feasible switching times within the horizon to determine the optimum sequence. Valid switching times are constrained by minimum and maximum phase durations. The problem can be re-formulated as an alternative dynamic programming problem by re-defining the control variable x_j to denote the amount of green plus yellow time allocated to stage s_i . The stages, in this case, correspond to the phase lengths during the horizon. The recursive optimization functional (forward DP), is:

$$f_j(s_j) = \text{Min}_{x_j} \left\{ R_j(s_j, x_j) + f_{j-1}^*(s_{j-1}) \right\}. \quad (13)$$

The return (or Performance Index) is now

$$R = \sum_a \int_{s_{j-1}}^{s_j} A^a(t) dt + \sum_a N_s^a. \quad (14)$$

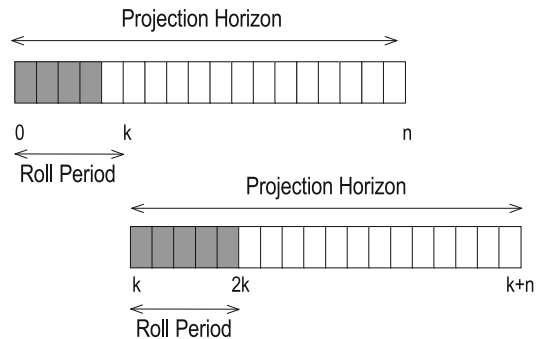
This calculation results in the optimal sequence of signal phase (stage) lengths during the horizon. By reformulating the DP model, computation is considerably more economical than in *OPAC I*. *OPAC II* lends itself more readily to operation in real-time than does *OPAC I*; however, it still requires information on arrivals (flows) over the entire horizon length.

OPAC III: A Rolling Horizon Approach

A typical horizon length is 1–2 min long. Obtaining accurate arrival predictions for this length of time is not feasible with current technology. To use only readily available flow data without degrading the performance of the optimization procedure, a 'rolling horizon' strategy is applied to the *OPAC II* algorithm. In this version, the horizon length, or the *Projection Horizon* is the period for which traffic patterns are projected and optimum phase change information is calculated. The key feature is that real-time data are required for only a small portion of the horizon.

Figure 17 is an illustration of the rolling horizon procedure. From detectors placed upstream of each approach actual arrival data for k intervals can be obtained for the beginning, or head, portion of the horizon. For the remaining $n - k$ intervals, the tail of the horizon, flow data may be obtained from a model. A simple model consists of a moving average of all previous arrivals on the approach. An optimal switching policy is calculated for the entire horizon, but only those changes which occur within the head portion are actually being implemented. In this way, *OPAC III* can dynamically revise the switching decisions as more recent (i.e., more accurate) real-time data continuously become available.

By placing the detectors well upstream of the intersection (10–15 s travel time) one can obtain actual arrival information for the head period. This allows for a more reliable estimation of delay for any given phase change decision. At



Optimization and Control of Urban Traffic Networks,
Fig. 17 Implementation of the rolling horizon approach in OPAC

the conclusion of the current head period, a new projection horizon containing new head and tail periods is defined with the new horizon beginning at (rolled to) the termination of the old head period. Calculations are then repeated for the new projection horizon. The roll period can be any multiple number of steps, including one. A shorter roll period implies more frequent calculations and, generally, closer to optimum (i.e., ideal) results. Extensive simulation and field tests have demonstrated the effectiveness of this strategy in achieving performance that approaches the theoretical optimum established by *OPAC I*. The placing of the detectors and the information flow among the intersection controllers is illustrated in Fig. 18.

OPAC IV: The Virtual Fixed Cycle Strategy

Rhythmic, or cyclic operation is essential for coordination of signals. Whereas an independent *opac* controller is cycle-free, linking of adaptive signals in a network configuration requires coordinated operation to facilitate opportunities for unimpeded progression. This is achieved in the *vfc-opac* model by re-introducing the concepts of cycle time and offsets in a benign manner that allows for increased flexibility of signal timing selection and control strategy implementation. *Vfc-opac* controlled intersections interact with neighboring intersections (fixedtime, or other *vfc-opac* controlled) in response to projected arrival flows from upstream feeder signals. The model offers, at the option of the user, a coordination-synchronization strategy that is suitable for implementation in arterials and in networks. The strategy is referred to as *virtualfixed-cycle* because from cycle to cycle the yield point, or local cycle reference point, is allowed to range about the fixed yield points dictated by the virtual cycle length and the virtual offset. This allows the synchronization phases to terminate early or extend later to better manage dynamic traffic conditions. *Vfc-opac* consists of a three-layer control architecture as shown in Fig. 19.

Layer 1: The *Local Control Layer* implements the *OPAC III* rolling horizon procedure using the dynamic programming model of *OPAC II*. It continuously calculates optimal switching

sequences for the Projection Horizon, subject to the *VFC* constraint communicated from Layer 3.

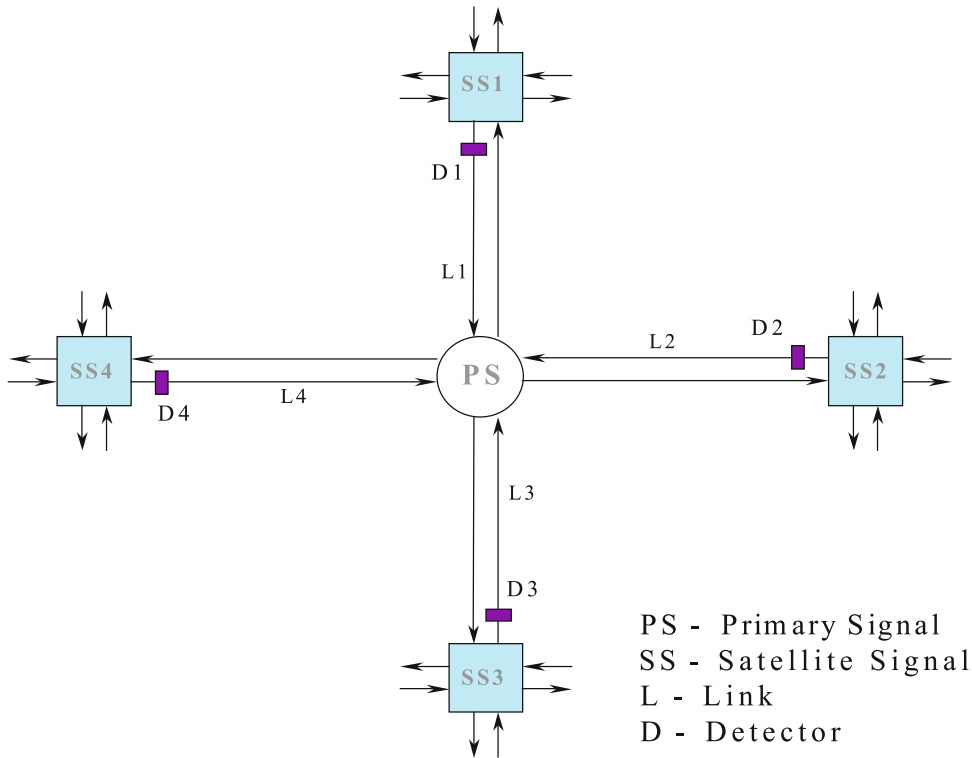
Layer 2: The *Coordination Layer* optimizes the offsets at each intersection (once per cycle). This is done by searching for the best offset of the *PS* (primary signal) within the mini-network shown in Fig. 4. A choice of three offset increments is being considered: 0 (no change), +2 s (move right one interval), −2-s (move left one interval). The cyclic flow profile associated with each incoming link is being discharged by the model through the intersection and projected to the downstream intersections, *SS* (satellite signals). All other parameters are being kept at their latest values in a relaxation mode. Since the coordination process is carried out in a distributed fashion at each intersection, each *SS*, in its turn, is also considered a *PS* of its own mini-network once during each cycle.

Layer 3: The *Synchronization Layer* calculates the network-wide *virtual-fixed-cycle* (once every few minutes, as specified by the user) in order to maintain a rhythmic operation of the signals in the network. The objective is to provide maximum leeway of phase switching timings as dictated by local conditions, yet maintain a capability for coordination with neighboring intersections by maintaining synchronicity of the signals which are linked in the network. The *virtual-fixed-cycle* (VFC) is calculated in a way that provides sufficient capacity at the most heavily loaded intersection while, at the same time, maintaining suitable progression opportunities among adjacent intersections. The VFC is calculated as follows:

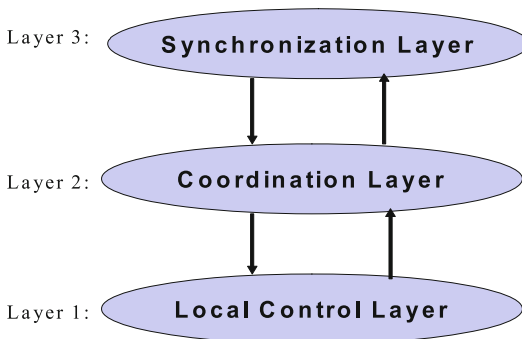
- Check if pre-set time period (3–5 min), or number of cycles counted, has elapsed
- Identify the dominant intersection (based on flow/saturation flow ratios)
- Establish bounds on VFC to satisfy the following requirements:

(a) Provide sufficient capacity

$$C_a \geq \sum_j (G_j)_{\min} + L$$



Optimization and Control of Urban Traffic Networks, Fig. 18 Information processing at an OPAC controlled intersection



$$C_b \geq \frac{k_m L}{k_m - \sum_j v_i}$$

(c) Do not exceed upper phase length limits

$$C_c \leq \sum_j (G_j)_{\max} + L$$

where, G_j = length of (green) phase j

Optimization and Control of Urban Traffic Networks, Fig. 19 Control architecture in VFC-OPAC

- $y_j = (q/s)_j$ = ratio of flow/sat-flow on phase j
- L = total lost time at the intersection

(b) Keep the degree of saturation under a preset maximum k_m (e. g., $k_m = 0.90$)

The virtual-fixed-cycle is chosen as the lowest value satisfying these requirements;

$$\text{VFC} = \min(C_a, C_b, C_c).$$

In addition, there may be exogenously determined limits on the cycle time

$$C_{\min} \leq VFC \leq C_{\max}.$$

The three-layer architecture is an effective means for implementing the distributed dynamic programming (DDB) algorithm. The VFC can be calculated separately for groups of intersections, as desired. The position of the marker line also determines the “virtual offset” of the signal. Over time the flexible cycle length and offsets are updated as the system adapts to changing traffic conditions. Due to the embedded flexibility this approach can provide improved local adaptability while maintaining network coordinability. Further information on the virtualfixed-cycle model is given in (Gartner et al. 2002).

Multi-Level Traffic Control Strategies

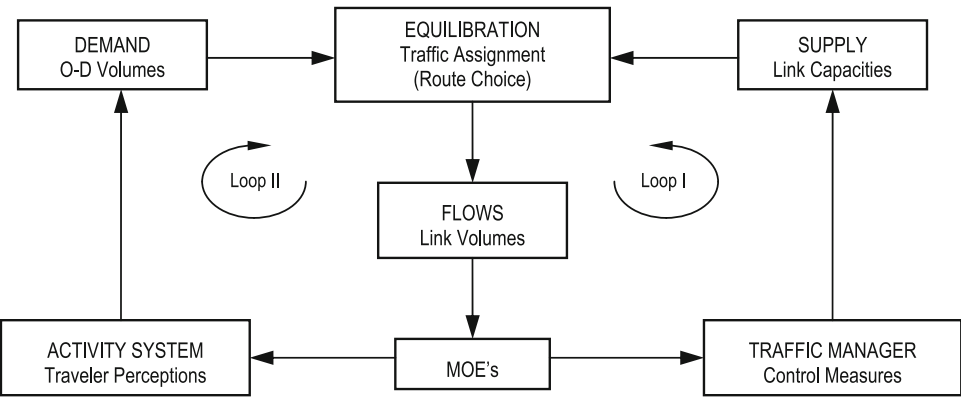
The performance of the transportation system is the result of a complex interaction between the physical supply of transportation facilities, and the individual trip-makers’ decisions. The supply includes the management policies and operational controls that are applied in these facilities. The flows in a transportation network are the result of an “equilibration” process between the demand for transport services and the supply of transport capacity. This was the basis for the development of “static” models for the design and planning of TSM actions (Gartner et al. 1980; Gershwin et al. 1978).

The extension of the static framework to the quasidynamic and dynamic cases, where on-line

control and guidance is being provided in response to real-time traffic information. The application of advanced technologies in sensing, communications and computation in intelligent transportation systems, coupled with advanced modeling concepts, provides great opportunities to improve the performance of traffic networks under both recurrent and non-recurrent conditions. Current research and development efforts are directed toward development of a Dynamic Traffic Assignment (DTA) capability to predict future traffic conditions and a Real Time Traffic Adaptive Control System (RT-TRACS) for the generation of signal control strategies. Although these models are intimately connected, their development has proceeded independently of each other so far. In the framework described herein the two models are integrated into a combined system.

The Traffic Management Problem

The traffic management problem can be viewed as the interaction between the urban traffic manager and the individual trip-maker and their (sometimes differing) perceptions of the performance of, or the supply provided by, the transportation system. A schematic illustration of the interactions in a transportation network is shown in Fig. 20. The physical transportation system and the socioeconomic activity system interact, via the equilibration process, to produce a set of flows on the links of the network. Two major feedback loops affect this process. In Loop I, the traffic manager assesses the system’s performance



Optimization and Control of Urban Traffic Networks, Fig. 20 Static interactions in a transportation system

according to his measures-of-effectiveness and intervenes in the physical transportation system to achieve his desired objectives (or, more accurately, the objectives of the community which he represents), such as reducing total travel times, minimizing pollution, increasing safety, etc. The trip-makers, on the other hand, assess the flows according to their own perceptions, which may be different from those of the traffic manager, and propagate an adjustment in the travel demand pattern via Loop II. Performance of the system is a result of the combined interaction of the demand and the supply feedback loops.

The problem can be set in the following context: given (fixed) demands for travel in an urban area and a (fixed) supply of transportation facilities, the traffic manager must consider a variety of management strategies to induce a traffic flow pattern that will meet, in an optimal way, the overall objectives of the community. The manager's criteria for evaluating his actions may include public interest measures of performance (MOP) such as travel time, energy consumption, noise and pollutant emissions, etc. On the other hand, individual trip-makers are assumed to minimize only their own travel costs (usually the travel time), subject to the constraints imposed by the physical system supply, the manager's actions, and the interactions resulting from the other trip-makers' decisions. The problem can be cast as *system-optimization* (the manager's objective) for flows that result from *user-optimization* (the trip-makers' objective). This leads to a compound mathematical optimization formulation, as follows:

$$\min_M \left\{ Z_1 \left[M, \arg \min_F Z_2(F) \right] \right\} \quad (15)$$

subject to flow conservation constraints:

$$\sum_{p \in P_{(i,k)}} h_p = r_{(i,k)} \quad \text{with } (i,k) \in I \quad (16a)$$

and non-negativity constraints:

$$h_p, r_{(i,k)} \geq 0 \quad (16b)$$

where:

M = the set of management variables under the control of the traffic manager

I = set of all OD pairs (i, k)

$P_{(i,k)}$ = set of all simple paths between OD pair (i, k)

f = set of all link flows

f_j = flow on link j

$r_{(i,k)}$ = rate of trip interchange (demand in vehicles) between origin node i and destination node k , with $(i, k) \in I$

h_p = flow on path p , with $p \in P_{(i,k)}$.

The compound objective function in Eq. (15) consists of the following components:

$$Z_1 = \sum_j f_j \cdot w_j(f_j, M) \quad (17a)$$

$$Z_2 = \sum_j \int_0^{f_j} c_j(x) \cdot dx \quad (17b)$$

where $c_j(f_j)$ is the average user-perceived travel cost function on link j and $w_j(f_j, M)$ is a more general performance function reflecting the multiplicity of objectives pursued by the traffic manager in the public's interest. The traffic management problem stated by Eq. (15) consists of a primary optimization program (the *system-optimization*), and a secondary optimization program (the *user-optimization*). The "argmin" of a mathematical program is the optimal solution of the program; therefore, $Z_1(\cdot)$ has to be evaluated at the optimal solution of the secondary optimization program and represents aggregate traffic performance in the network with user-optimal flows.

An alternative mathematical representation of the traffic management problem is given by Tan et al. (1979) and is labeled "hybrid optimization." In this case the objective function of the traffic management problem coincides with the system optimization objective described in Eq. (17a), while the user optimization objective acts as a constraint. Tan et al. show that Wardrop's First Principle describing the user optimization problem can be fully expressed mathematically as follows:

1. Every path p must carry non-negative flows:

$$h_p \geq 0$$

2. Path flows of the same O-D pair sum to the required flow $\sum_{p \in P_{(i,k)}} h_p = r_{(i,k)}$ with $(i,k) \in I$
3. Utilized paths have the same cost c_q which is equal to the minimum cost and paths with a greater cost than the minimum carry no flow. This is expressed by the following:

$$c_q(h_q, w_q) \geq \sum_{p \in P_{(i,k)}} h_p \cdot c_p(h_p, w_p) \cdot c_p(h_p, w_p) \\ \times / r_{(i,k)} \text{ with } (i,k) \in I \text{ and } p \in P_{(i,k)}.$$

Thus the hybrid optimization formulation of the traffic management problem is:

$$\text{Minimize } Z_1 = \sum_j f_j \cdot w_j(f_j, M)$$

subject to conditions (1.), (2.) and (3.) above.

Decision variables in the primary program are the set of management variables M under the control of the traffic manager. Decision variables in the secondary program are the link flows F , given the OD demands $r_{(i,k)}$. Major categories of actions that can be considered are: (a) actions to ensure the efficient use of existing road space; (b) actions to reduce vehicle use in congested areas; and (c) actions to improve transit service. For example, representative variables that are particularly amenable to formal optimization include traffic signalization variables (cycle time, green splits, offsets, signal phasing and phase sequencing) traffic operations variables (e. g., channelization, oneway streets, metering, variable message signs, reversible lanes) and preferential treatment variables (e. g., reserved or preferential lanes, exclusive transit lanes, transit vehicle preemption of traffic signals, special turning lanes, truck routes, parking control). The compound formulation and the hybrid-optimization are similar models; however, the latter has been more thoroughly investigated and, therefore, is more amenable to computation.

The framework given above for the static traffic management problem can be extended to the

dynamic case. In principle, the same model applies; however, both the OD demands and the control actions M are now time-dependent. Further information is given in (Gartner and Stamatiadis 1998).

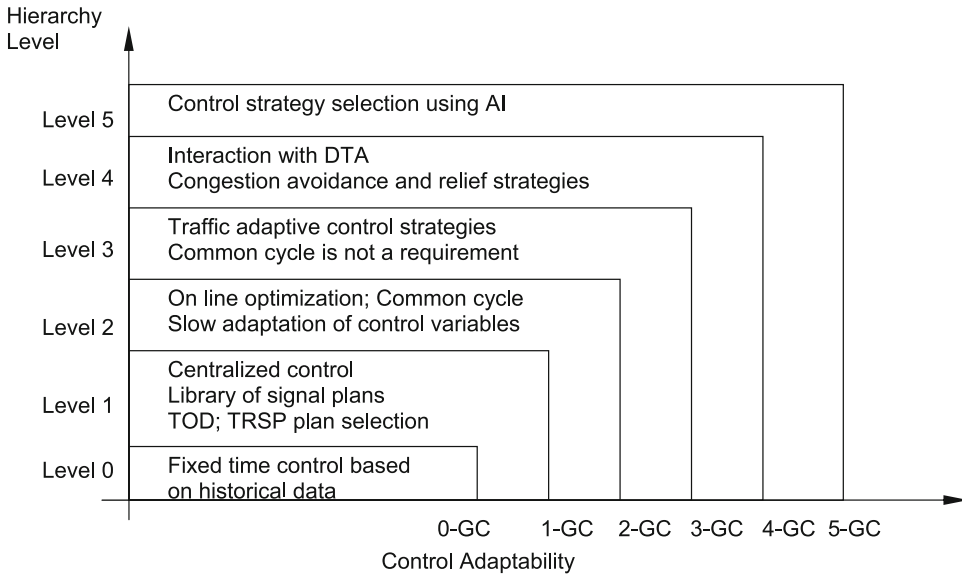
Signal Control and Traffic Assignment

Although signal control and route choice are intimately intertwined on the street, there is little practical experience with the integration of signal control and traffic assignment from a modeling standpoint. As indicated previously, most modeling efforts were done independently, i.e., assignment separate from control.

Smith (1980) proposed a combined assignment-control model to take into consideration the effect of signal settings on route choice. The model is based on the P_0 control policy which is designed to maximize the “travel capacity” of the road network, under the assumption that all drivers seek their own best route. From computations conducted by Ghali and Smith (1994) it was concluded that the P_0 control policy performs overall better than two standard control policies, (a) the local delay minimization for current traffic patterns at each intersection and (b) Webster’s equisaturation method, where green times are selected to equalize the degree of saturation on competing approaches. Gartner and Al-Malik (1996) presented a combined network model which simultaneously accounts for both the route choices made by motorists and the desired signal controls to match these choices. An important feature of this approach is that offset optimization can also be incorporated.

Real Time Traffic-Adaptive Control System (RT-TRACS)

RT-TRACS consists of a multi-level hierarchy of traffic control strategies and is illustrated in Fig. 21. This system provides a modern approach to adaptive traffic signal control and has been specifically designed to facilitate integration with traffic prediction (Gartner et al. 1995). The section below describes an RT-TRACS version that can also be incorporated within the framework of a dynamic traffic management system which is a dynamic version of the system illustrated in Fig. 20.



Optimization and Control of Urban Traffic Networks, Fig. 21 Nested hierarchy of RT-TRACS control levels

The degree of adaptability corresponds to the generation of the control strategy development, where n -GC is the n th generation control. Each level encompasses the capabilities of the lower levels in a nested fashion. The different levels and their interaction with predicted traffic data are described below.

Level 0 is the most basic type of signal control. The traffic engineer determines timing plans manually, or by some PC-based optimization program, using historical volume data. This level is suitable for recurrent traffic patterns, i.e., when there is little need to predict traffic flows, and there is infrequent updating of plans.

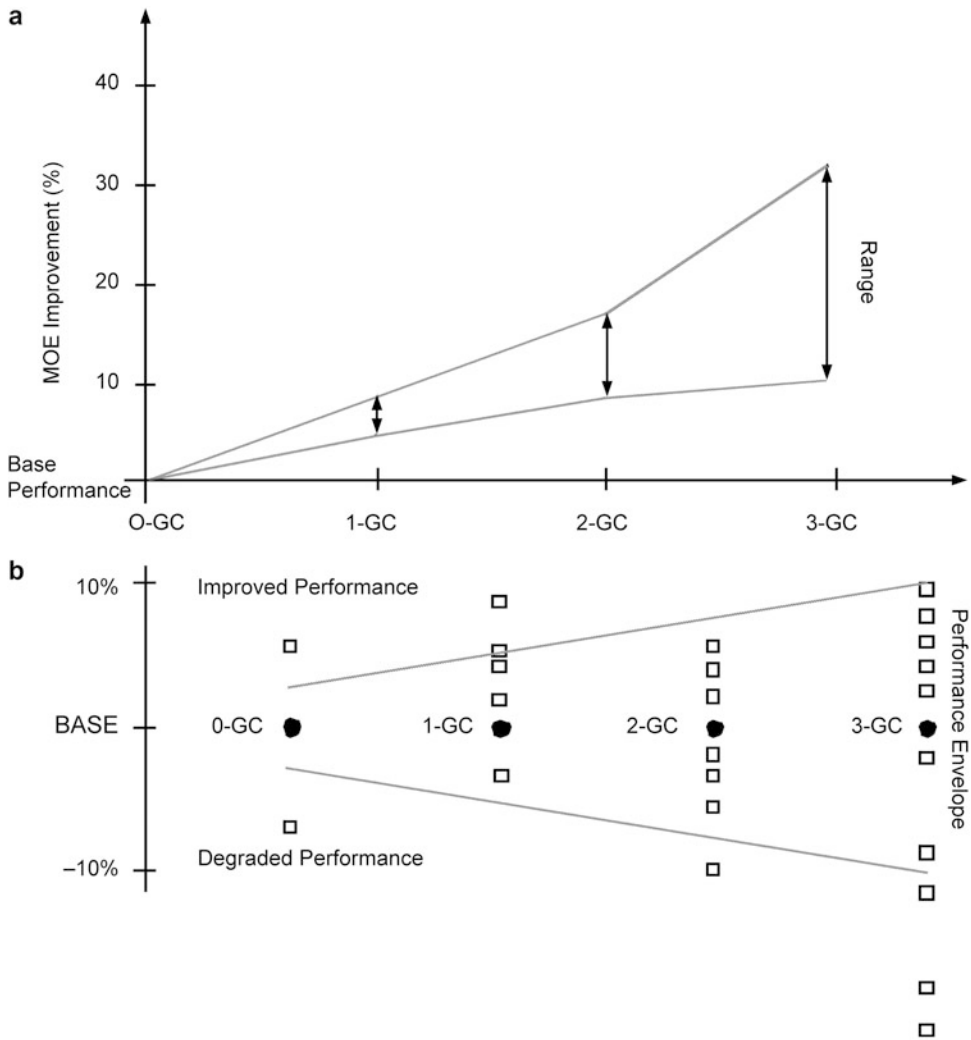
Level 1 corresponds to the 1-GC system: centralized control with limited availability of traffic surveillance and communications. Timing plans are calculated off-line using a multitude of scenarios to create a library of plans that can be activated in response to changing traffic and network conditions. Time of day (TOD) and traffic responsive (TRSP) options are available. Update interval is 15 min or higher. This level is suitable for application in the quasi-dynamic case. Integration with DTA is achieved in the following way: given a predicted O-D matrix, the system selects the most suitable timing plan in the library for the

upcoming period. The plan should be created by a combined traffic assignment/signal optimization procedure and should match the current demand and supply conditions.

Level 2 (2-GC) is centralized control with on-line optimization using a fixed, common cycle. Similar to Level 1, time frames for optimization are shorter (5 to 10 min), plans are being generated on-line. This level of control can respond more rapidly to changes in travel demands and/or capacities. For example, progression schemes for priority routes can be generated on-line in conjunction with the traffic prediction module.

Level 3 (3-GC) is a fully-adaptive traffic signal control system that may also incorporate the capabilities of the previous levels if warranted. A common, fixed cycle time is not required for coordination. Capability of phase sequence optimization is available at selected locations. Example of a 3-GC strategy: the OPAC network version. This level of control is capable of reacting to rapidly changing traffic conditions and can also be used in conjunction with assignment capabilities to provide optimal control and rerouting of traffic. The latter capability is incorporated in the next level.

Level 4 includes all the capabilities of 3-GC with additional intelligence so that the system can interact with the DTA module. It can provide



Optimization and Control of Urban Traffic Networks, Fig. 22 (a) Expected relative performance of control generation. (b) Reported relative performance of control generation

dynamic priority control on selected routes and implement congestion avoidance and relief strategies. This level is most suitable for handling incident and accident conditions.

Level 5 is a super level that makes the most efficient use of the available array of control strategies based on accumulated experience under local conditions. Selection of the appropriate control strategy for a particular condition can be invoked by Artificial Intelligence technology. Practical experience has shown that no single strategy is optimal for the entire spectrum of demand and supply variations.

Therefore, it is beneficial to have available an arsenal of strategies and to be able to choose the most appropriate one when needed.

RT-TRACS provides a new design that is different from the traditional signal control generation strategies that are still state-of-the-art. In the traditional strategies no consideration is given to re-assignment of traffic, whereas here assignment is an integral part of the signal control generation process. The mutual interdependence between route choice and signal control is embodied at all levels of the hierarchy.

Control System Performance – Analysis of Complexity

The hierarchical framework described above offers the possibility to tailor the system's capabilities to particular needs and means. Advanced strategies can be deployed gradually and can coexist with lower level strategies. The full spectrum of capabilities can be built-up gradually in terms of deployment of sensors, surveillance equipment, controllers and communications, as well as central control hardware and software. It has been common wisdom that increasing responsiveness will contribute to improved traffic performance. Experience indicates that this has not always been the case. The more complex a system is the wider is the dispersion of the performance results (Gartner et al. 1995). The argument can be best illustrated with reference to Fig. 22. It shows the spreads in performance that are being perceived to exist among the different control generations (Fig. 22a), as well as the spreads in performance that were actually measured in the field (Fig. 22b). The points shown in the figure were collected from published reports in the literature. These figures illustrate an interesting phenomenon: the more "advanced" (i.e., responsive) strategies do not lead to improved performance all the time, notwithstanding the common perception that they do. More likely, they lead to a wider spread in performance results compared with a common basis. In some cases, due to reasons that are not completely explained, traditional off-line methods perform better than responsive methods. Therefore, one of the principal objectives of any intelligent control system would be to invoke a particular control strategy that will be most suitable for existing conditions so that overall performance of the system is optimized.

An expert system can be used to make sure that we implement a particular generation of control strategies when it is likely to yield the best performance. Development of the system requires careful characterization of signal networks and identification of the particular traffic flow patterns that are most amenable to benefit from a particular control strategy. The underlying proposition is

that lower level strategies may often be as good, or better than higher level, more advanced strategies. By recognizing the conditions under which the particular strategies perform best, we can optimize overall system performance. This is the task of the 5th generation control level.

Future Directions

With the continuing developments in technology, one can expect to have gradual improvements in the technology of detection, communication and optimization which would lead to better real-time information on the status of the transportation system and, thus, to improved online control and performance. Specific areas in which advances can be sought and are likely to be achieved are: traffic-adaptive control of signal systems and the combined optimization of controls and routing. In particular, the latter would benefit from the proliferation of in-vehicular route guidance devices that are tied-in to advanced traveler information systems.

Bibliography

Primary Literature

- Bellman R, Dreyfus S (1962) Applied dynamic programming. Princeton University Press, Princeton
- Gartner NH (1972) Constraining relations among offsets in synchronized signal networks. *Transp Sci* 6:88–93
- Gartner NH (1983) OPAC: a demand-responsive strategy for traffic signal control. *Transp Res Rec* 906:75–81
- Gartner NH, Al-Malik M (1996) A combined model for signal control and route choice in urban networks. *Transp Res Rec* 1554:27–35
- Gartner NH, Stamatiadis C (1998) Integration of dynamic traffic assignment with real-time traffic adaptive signal control. *Transp Res Rec* 1644:150–156
- Gartner NH, Stamatiadis C (2002) Arterial-based control of traffic flow in urban grid networks. *Math Comput Model* 35:657–671
- Gartner NH, Gershwin SB, Little JDC, Ross P (1980) Pilot study of computer-based urban traffic management. *Transp Res* 14(1/2):203–217
- Gartner NH, Assmann SF, Lasaga F, Hou DL (1991) A multiband approach to arterial traffic signal optimization. *Transp Res* 25B(1):55

- Gartner NH, Stamatiadis C, Tarnoff PJ (1995) Development of advanced traffic signal control strategies for ITS: a multi-level design. *Transp Res Rec* 1494:98–105
- Gartner NH, Pooran FJ, Andrews CM (2002) Optimized policy for adaptive control strategy in real-time adaptive control systems: Implementation and field testing. In: *Transp Res Rec*, vol 1811. Transportation Research Board, Washington, DC, pp 148–156
- Gerlough DL, Huber MJ (1975) Traffic flow theory. Special report 165. Transportation Research Board, Washington, DC
- Gershwin SB, Ross P, Gartner NH, Little JDC (1978) Optimization of large traffic systems. *Transp Res Rec* 682:8–15
- Ghali MO, Smith MJ (1994) Comparisons of the performances of three responsive traffic control policies, taking drivers' day-to-day route choices into account. *Traffic Eng Control* 1994:555–560
- Grafton RB, Newell GF (1967) Optimal policies for the control of an undersaturated intersection. In: Edie LC, Herman R, Rothery R (eds) *Vehicular traffic science*. Elsevier, New York, pp 239–257
- Greenshields BD, Schapiro D, Ericksen EL (1947) Traffic performance at urban street intersections. Technical report, no 1. Yale Bureau of Highway Traffic, New Haven
- Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT – A traffic responsive method of coordinating signals. TRRL report no LR 1014. Transportation and Road Research Laboratory, Crowthorne
- Little JDC (1966) The synchronizing of traffic signals by mixed integer linear programming. *Oper Res* 14:568–594
- Little JD, Kelson MD, Gartner NH (1982) MAXBAND: a program for setting signals on arteries and triangular networks. *Transp Res Rec* 795:40–46
- Morgan JT, Little JDC (1964) Synchronizing traffic signals for maximal bandwidth. *Oper Res* 12:896–912
- Robertson DI (1969) TRANSYT: a traffic network study tool. TRRL report no LR 253. Transportation and Road Research Laboratory, Crowthorne
- Smith MJ (1980) A local traffic control policy which automatically maximizes the overall travel capacity of an urban network. *Traffic Eng Control* 1980:298–302
- Tan HN, Gershwin SB, Athans M (1979) Hybrid optimization in urban traffic networks. Report No DOT-TSC-RSPA-79-7. Transportation Systems Center, US Dept of Transportation, Washington, DC
- Webster FV (1958) Traffic signal settings. HMSO, London

Books and Reviews

- Gartner NH, Improta G (eds) (1995) *Urban traffic networks: dynamic flow modeling and control*. Springer, Berlin
- Gordon RL et al (1996) *Traffic control systems handbook*. Federal highway administration. Report FHWA-SA-95-032. US Dept of Transportation, Washington, DC
- Homburger WS (ed) (1982) *Transportation and traffic engineering handbook*, 2nd edn. Prentice-Hall, Englewood Cliffs
- Kerner BS (2004) *The physics of traffic*. Springer, Berlin
- Organization for economic cooperation and development (1987) *Dynamic traffic management in urban and suburban road systems*. OECD, Paris
- Papageorgiou M (ed) (1991) *Concise encyclopedia of traffic & transportation systems*. Pergamon Press, Oxford
- RiLSA (2002) *Richtlinien für Lichtsignalanlagen für den Strassenverkehr*. In: Köln. Forschungsgesellschaft für Strassen-und Verkehrswesen

Freeway Traffic Management and Control

Andreas Hegyi¹, Tom Bellemans² and Bart De Schutter³

¹Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands

²Hasselt University, Diepenbeek, Belgium

³Delft Center for Systems and Control, Delft University of Technology, Delft, The Netherlands

Article Outline

Glossary

Definition of the Subject

Introduction

Sensor Technologies

Traffic Flow Modeling

Freeway Traffic Control Measures

Network-Oriented Traffic Control Systems

Future Directions

Conclusion

Bibliography

Abbreviations

ADAS	Advanced driver assistance systems
AHS	Automated highway system
IVHS	Intelligent vehicle/highway system
MPC	Model predictive control
OD	Origin-destination

Glossary

Feedback control In Fig. 1 the feedback control structure is shown (Åström and Wittenmark 1997). In contrast to the feedforward control structure, here the behavior of the outputs is coupled back to the controller (hence the name feedback). This structure is also often referred to as “closed-loop” control. The main advantages of a feedback controller over a feedforward controller are that (1) it may have a quicker response

(resulting in better performance), (2) it may correct undesired offsets in the output, (3) it may suppress unmeasurable disturbances that are observable through the output only, and (4) it may stabilize an unstable system.

Feedforward control The block diagram of a feedforward control structure is shown in Fig. 2 (Åström and Wittenmark 1997). The behavior of process **P** can be influenced by the control inputs. As a result the outputs (measurements or observations) show a given behavior. The controller **C** determines the control inputs in order to reach a given desired behavior of the outputs, taking into account the disturbances that act on the process. In the feedforward structure the controller **C** translates the desired behavior and the measured disturbances into control actions for the process. The term feedforward refers to the fact that the direction of the information flow in the system contains no loops, i.e., it propagates only “forward.” The main advantages of a feedforward controller are that the complete system is stable if the controller and the process are stable and that its design is in general simple.

Model predictive control Model predictive control (MPC) is an extension of the optimal control framework (Camacho and Bordons 1995; Maciejowski 2002). In Fig. 3 the block diagram of MPC is shown. In MPC, at each time step k the optimal control signal is computed (by numerical optimization) over a prediction horizon of N_p steps. A control horizon N_c ($< N_p$) can be selected to reduce the number of variables and to improve the stability of the system. Beyond the control horizon the control signal is usually taken to be constant. From the resulting optimal control signal only the first sample of the computed control signal is applied to the process. In the next time step $k + 1$, a new optimization is performed with a prediction horizon that is shifted one time step ahead, and of the resulting control signal again only the first sample is applied, and so on. This scheme, called rolling horizon, allows for

updating the state from measurements, or even for updating the model in every iteration step. In other words, MPC is equivalent to optimal control extended with feedback. The advantage of updating the state through feedback is that this results in a controller that has a low sensitivity to prediction errors. Regularly updating the prediction model results in an adaptive control system, which could be useful in situations where the model significantly changes, such as in case of incidents or changing weather conditions.

Optimal control Optimal control is a control methodology that formulates a control problem in terms of a *performance function*, also called an *objective function* (Lewis 1992). This function expresses the performance of the system over a given period of time, and the goal of the controller is to find the control signals that result in optimal performance. Depending on the mathematical description of the control problem, there exist several methods for the optimization of the control input including analytic and numerical approaches. Optimal control can be considered as a feedforward control approach.

more reliable travel times, or reduced emissions and noise production.

The tools used for this purpose are in general changeable signs (including traffic signals, dynamic speed limit signs, and changeable message signs), radio broadcast messages, or human traffic controllers at the location of interest. Moreover, currently the possibilities of assisting, informing, and guiding drivers via in-car systems are also being explored.

The term *dynamic traffic management* includes besides dynamic traffic control also the management of emergency services and nonautomated procedures (such as the implementation of predefined traffic control scenarios during special events), typically performed in traffic control centers. However, in this entry the focus is on automatic control methods. Furthermore, this entry deals exclusively with dynamic freeway traffic control techniques. Given the differences in traffic operation (e.g., higher speed limits) and in traffic infrastructure (e.g., intersections versus on-ramps and off-ramps), the control measures that can be implemented for urban and for freeway traffic differ. The interested reader is referred to Daganzo (1997), Gartner and Improta (1995), and Papageorgiou (1991) for an overview of urban traffic control.

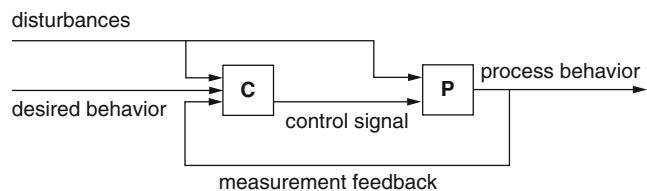
Definition of the Subject

The goal of this entry is to provide an overview of dynamic traffic control techniques described in the literature and applied in practice. *Dynamic traffic control* is the term to indicate a collection of tools, procedures, and methods that are used to intervene in traffic in order to improve the traffic flow on the short term, i.e., ranging from minutes to hours. The nature of the improvement may include increased safety, higher traffic flows, shorter travel times, more stable traffic flows,

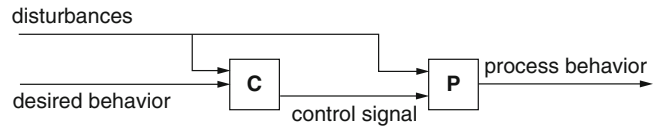
Introduction

The number of vehicles and the need for transportation are continuously growing, and nowadays cities around the world face serious traffic congestion problems: Almost every weekday morning and evening during rush hours the capacity of many main roads is exceeded. Traffic jams do not only cause considerable costs due to unproductive time losses; they also augment the

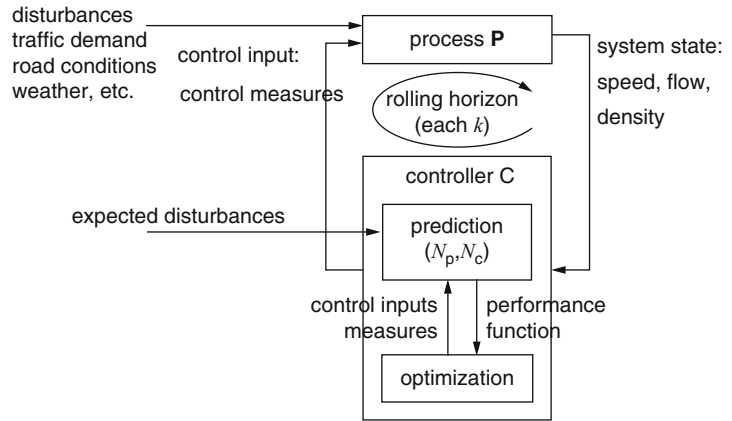
Freeway Traffic Management and Control, Fig. 1 The feedback control structure



Freeway Traffic Management and Control, Fig. 2 The feedforward control structure



Freeway Traffic Management and Control, Fig. 3 The model predictive control (MPC) structure



probability of accidents and have a negative impact on the environment (air pollution, lost fuel) and on the quality of life (health problems, noise, stress).

One solution to the ever-growing traffic congestion problem is to extend the road network. Extending the freeway infrastructure is rather expensive, and in many countries this option is currently not considered to be a viable solution. Moreover, in densely populated areas building new roads is sometimes even unfeasible due to spacial limitations. Furthermore, there are often also other socioeconomic objectives to be achieved, such as environmental objectives, which are considered alongside the objective of reducing congestion. Dynamic traffic control is an alternative that aims at increasing the safety and efficiency of the existing traffic networks without the necessity of adding new road infrastructure.

Since the 1960s traffic control is applied on freeway systems. However, during the last decades there have been developments in traffic science, traffic technology, control theory, and in the typical traffic patterns that all have consequences for the

most appropriate traffic control approach. These developments will be discussed in the next sections.

The Need for Network-Oriented Automatic Traffic Control: Developments

The increasing complexity of the congested traffic patterns and the increasing availability of traffic control measures motivate the increasing usage of automatic traffic control systems and the increasing interest in network-oriented control over the last decades. The interest in network-oriented control from the practitioners' point of view is also motivated from policies aiming at socioeconomic goals, such as the efficient transport over important network corridors.

The fact that the length, the duration, and the number of traffic jams continues to grow has consequences for traffic control. When there are more locations with congestion, the available control measures have to solve more problems, which imply a higher complexity. Since nowadays the chances are higher that a vehicle encounters more than one traffic jam during one trip, the traffic control measures influencing a vehicle in one

traffic jam will also influence the other jam(s) that the vehicle encounters. Therefore, the spatial interrelations between traffic situations at different locations in the network get stronger, and consequently the interrelations between the traffic control measures at different locations in the network also get stronger. These interrelations may differ per situation (and depend on, e.g., network topology, traffic demand) and the control measures may even counteract each other. For the various traffic management agencies or local governments that are responsible for different parts of the traffic network, this means that there is a need for a stronger cooperation and agreement on how the common network goals should be achieved. Similarly, for the automatic control methods, coordinated control strategies are required in these cases, to ensure that all available control measures serve the same objective, or at least that they do not counteract each other.

Another development is that freeways are equipped with more and more traffic control measures. The increasing number of control measures augments the controllability of the freeways. However, with this development the number of possible combinations of control measures also increases drastically, which in its turn increases the complexity of the dynamic traffic management problem.

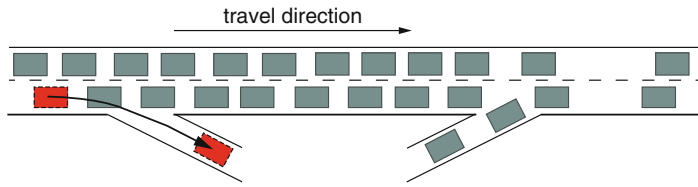
On modern freeways often a large amount of data is available on-line and off-line. This data can serve as a basis for choices about appropriate control measures given the actual and expected traffic situation. However, the available data is currently not fully utilized either by traffic control center operators, whose actions are typically based on heuristic reasoning, or by automatic control measures, which mostly use local data only. Traffic data may also contain information about the current disturbances of the network (incidents, weather influences, unexpected demands) and information about the traffic system at a network level (about route choice and origin-destination relationships). The origin-destination (OD) matrix describes the traffic demand (vehicles per hour) appearing at each origin in a traffic network toward each destination in the network. An OD matrix may be time-varying and can be calculated at different levels of temporal

aggregation, e.g., hourly, peak or interpeak, 24 h. Methods have been developed to estimate such OD relationships from traffic measurements and to estimate the traffic state (e.g., speeds, flows, and densities) in networks that are incompletely equipped with detectors (Hegyi et al. 2006; Wang and Papageorgiou 2005). The methods can be used to supply a traffic control system with more accurate data, leading to better control actions.

These developments motivate the application of automatic control systems that can handle complex traffic scenarios, multiple control measures, and a large amount of data, and that can benefit from the network-oriented information by selecting appropriate control measures for given OD patterns and disturbances.

Regardless of these developments the three effects that cause the majority of suboptimal traffic network performances – in terms of travel time – have remained the same. Therefore, the primary goal of traffic control was and is to resolve the following effects/issues:

- **Bottlenecks.** Typical bottlenecks are freeway sections with an on-ramp, bridges, tunnels, curves, grades, and merge and diverge areas. The performance degradation typically originates from the phenomenon that the maximum achievable outflow from a traffic jam created at a bottleneck is often lower than the capacity of the road. This phenomenon is often called the capacity drop. A special case of a bottleneck is an upstream propagating jam that is growing at the tail by the incoming vehicles and resolving at the head by the leaving vehicles. A moving jam can be a serious bottleneck as it could reduce the maximum outflow to 70% of the capacity (Kerner 2002), while the capacities of the other bottlenecks are in the range of 85–100% (Hall and Agyemang-Duah 1991; Kerner 2002). Dynamic traffic control measures may help to prevent a traffic breakdown at a bottleneck or to improve the flow when a breakdown has occurred.
- **Suboptimal route choice.** In a dynamically changing network with jams, incidents, and road works, the driver may not always be informed sufficiently well to make the optimal

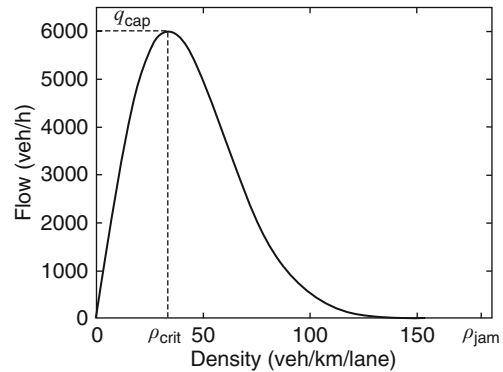


Freeway Traffic Management and Control, Fig. 4 Congestion caused by high on-ramp demand could also result in the blocking of an upstream off-ramp

route choice. Furthermore, even if each individual driver has found the quickest (or in general, least costly) route to his or her destination, it may not lead to optimal performance at the network level, as known from the famous example of Braess (1968). Systems that influence the drivers' route choice may contribute to a better performance for the users, the network, or even both.

- **Blocking.** The tail of a traffic jam on a freeway may propagate so far upstream that it blocks traffic on a route that is not leading over the bottleneck that has caused the jam. A typical case is when a traffic jam created on the freeway at an on-ramp propagates back to an upstream off-ramp and blocks the traffic that wants to leave via the off-ramp. Figure 4 illustrates a situation where off-ramp traffic is blocked by a jam originating from the downstream on-ramp. All control measures that can limit the length of a traffic jam may in principle be applied to prevent blocking.

Automatic traffic control strategies try to optimize traffic network performance. A simplified, idealized description of the operation of traffic in the network links is given by what is known in traffic theory as the fundamental diagram (May 1990). The fundamental diagram describes steady-state traffic operation on a homogeneous freeway (i.e., the spatial gradients of speed, flow, and density are equal to zero) as illustrated in Fig. 5. For low traffic densities, the relation between traffic density and traffic flow is nearly linear. For traffic densities smaller than the critical density ρ_{crit} , the traffic flow on the freeway increases with increasing traffic density (Fig. 5), despite the fact that the average speed decreases



Freeway Traffic Management and Control, Fig. 5 A flow-density fundamental diagram for a three-lane freeway. As long as the traffic density on the freeway is smaller than the critical density ρ_{crit} , the traffic flow on the freeway increases with increasing traffic density. If the traffic density reaches the critical density, the flow is maximal and equal to the freeway capacity q_{cap} . If the traffic density further increases, the traffic flow on the freeway starts to decrease with increasing traffic density until the traffic stalls at the jam density ρ_{jam}

with increasing traffic density. Traffic operation is in a stable regime for traffic densities lower than the critical density. The maximal flow that can be achieved on the freeway, the capacity q_{cap} , is reached for a traffic density equal to the critical density and the resulting average speed of the vehicles is called the critical speed. If the critical density is exceeded, the average speed continues to decrease and the traffic flow decreases with increasing density. For traffic densities higher than the critical density, congestion sets in and an unstable traffic regime results. Typical values of ρ_{crit} and q_{cap} for a three-lane highway are 33.5 vehicles per kilometer and per lane and 6000 vehicles per hour, respectively (Papageorgiou et al. 1990a).

Based on the discussion of the fundamental diagram presented above, it can be observed that automatic control strategies can try to prevent or to reduce congestion by steering the state of traffic operation toward the stable region of operation. Here the fundamental diagram is merely presented as a seminal approach to traffic state analysis. However, in the literature alternative approaches to traffic state analysis have been reported, e.g., Kerner (2004) has proposed the three phase theory, introducing the concepts “wide moving jams” and “synchronized traffic.” The work by Treiber et al. (2000) and Lee et al. (1999) has added additional traffic states to the analysis such as “oscillating congested traffic” and “homogeneous congested traffic.”

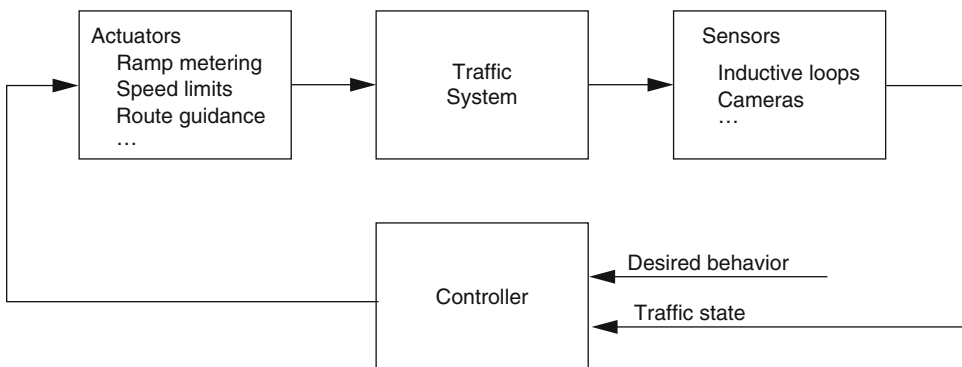
Automatic Traffic Control

Dynamic traffic management systems typically operate according to the feedback control concept known from control systems theory, as shown in Fig. 6. The traffic sensors provide information about the current traffic state, such as speed, flow, density, or occupancy. The controller determines appropriate control signals that are sent to the actuators (depending on the system the changes in the control signal may be implemented instantly or may need to be phased in). The reaction of the traffic system is measured by the sensors again, which closes the control loop. If the new

measurements show a deviation from the desired behavior (caused, e.g., by unforeseen disturbances), the new control signals are adapted accordingly. Note that there also exist traffic control systems that have a feedforward structure, e.g., the demand-capacity ramp metering approach that will be discussed in section “[Ramp Metering](#).”

We define an “appropriate control” signal in terms of a control objective. From the network operator’s point of view typical objectives are:

- **Efficiency.** Efficiency is often expressed in terms of throughput or travel time. This objective is shared by the network operators and the individual drivers. Nevertheless, situations may arise when the network operator and the individual driver have conflicting interests, e.g., minimizing the total travel time in a network (network optimum) may be conflicting with individually minimizing the travel times (user optimum). This will be discussed in more detail in section “[Route Guidance](#).”
- **Safety.** In traffic control, safety may be a direct goal of the control or it may be a constraint (boundary condition) that should be satisfied. For example, dynamic speed limits or variable message signs may reduce the speed limit or give a warning under adverse weather conditions or poor visibility conditions in order to improve safety. Other systems may have other primary goals, such as



Freeway Traffic Management and Control, Fig. 6 Schematic representation of the dynamic traffic management control loop. Based on the measurements provided by the sensors the controller determines the

control signals sent to the actuators. Since the control loop is closed, the deviations from the desired traffic system behavior are observed and appropriate control actions are taken

improving the flow, and in these cases the control systems are often still required to be safety-neutral compared with the situation without control. There may also be an interaction between safety and efficiency, which has to be taken into account in the design of the control system. This interaction may be related to the following three processes. First, a safer traffic system in general results in fewer accidents and therefore more often higher flows may occur. Second, if congestion is prevented by an appropriate control method, safety may be increased due to the more homogeneous flows. Third, lower speeds and densities in general positively influence safety. More specifically, Brownfield et al. (2003) observed that for freeway sites, the accident rate in congested conditions was nearly twice the rate in uncongested conditions. However, the proportion of accidents that were serious or fatal was lower in congested conditions than in uncongested conditions. Hence, depending on the network operator's definition of safety, safety and efficiency may be conflicting or nonconflicting objectives.

- **Network reliability.** Even if not every traffic jam can be prevented, it is valuable for drivers when the travel time to their destinations is predictable, since good arrival time estimations make departure time choices easier. Therefore, improving the network reliability/predictability serves the economic efficiency of the network and improves driver comfort. Traffic control in general can improve reliability (predictability) by aiming at the realization of predicted travel times, or the reverse, by predicting realizable travel times, or both. Furthermore, network reliability can be improved by measures that aim at synchronization of the traffic demand and the capacity supply of the network, and at a better distribution traffic of flows over the network. For a more elaborate discussion on network reliability, we refer the interested reader to Bell (2000), Chen et al. (2002), Lo et al. (2006), and Taylor (1999a, b).
- **Low fuel consumption, low air, and noise pollution.** In general, congestion contributes to less smooth journeys (more deceleration-

acceleration movements), which increases emissions. In or near urban areas the environmental effects of traffic may be considered more important than, efficiency, for example, which can result in a different trade-off between the two objectives. An example of such a trade-off is between travel speed and air pollution (André and Hammarström 2000). The typical measure for these purposes is speed limitation.

Another important aspect of a traffic control system are the constraints due to physical, technical, or policy-based limitations. Such constraints may include minimum and maximum ramp metering rates, maximum on-ramp queue length, minimum and maximum dynamic speed limit values, etc. The automatic traffic controller is required to cope with these constraints.

Entry Overview

The remainder of this entry is organized as follows. We discuss the sensor technologies used in the context of freeway traffic control in section “[Sensor Technologies](#).” In section “[Traffic Flow Modeling](#)” we address traffic flow models, which play an important role in the design and evaluation of traffic control strategies. Next, the most frequently used freeway control measures are discussed in section “[Freeway Traffic Control Measures](#).” While in section “[Freeway Traffic Control Measures](#)” the focus is on the individual control measures, in section “[Network-Oriented Traffic Control Systems](#)” we discuss the approaches to combine and to integrate several control measures for network-oriented control. We conclude in section “[Future Directions](#)” by considering the new developments that are expected to play a role in future freeway traffic control systems.

Sensor Technologies

In order to implement traffic responsive freeway control, traffic measurements need to be collected at different locations throughout the freeway network. This section first deals with the most common traffic variables and traffic sensors to collect

them. Next, the need for traffic demand and traffic routing data, and the way these data can be obtained using common traffic measurements, is briefly addressed. New data collection technologies that will play an important role in future freeway control systems are discussed in section “Future Directions.”

Measurements

Traditionally the following traffic variables are measured to determine the traffic state on a freeway (May 1990): the traffic flow or the traffic intensity on the freeway, the average speed of the vehicles, the traffic density, the occupancy level of the freeway, the time headways (and in some cases the distance headways), and the speed variance. Note that occupancy is defined as the relative time (percentage) that the traffic sensor is occupied by a vehicle. In practice, it is often used as a surrogate measure for traffic density since it is directly related to density (as long as the average vehicle length is constant) and can be measured more easily than density.

Depending on the application and on the traffic measurement system, several levels of detail can be distinguished. The traffic variables can either be measured for every freeway lane separately or a value averaged over all lanes of the freeway can be obtained. Some measurement systems allow for a classification of the vehicles in categories based on their size (e.g., trucks versus cars) and provide the traffic variables per category. Furthermore, instantaneous values of the measured traffic variables can be provided or values averaged over a time period. The period over which the measurements are averaged can range from seconds over minutes to hours. As a rule of thumb, one can assume that the higher the level of detail of the data collected, the higher the cost of the measurement system involved.

In real-life situations, the measurements that are provided by the traffic sensors will contain measurement errors. These errors include incidental missing values, incidental faulty measurements, biased measurements, and missing values over a period of time. Given the importance of the traffic measurements in the dynamic traffic management control loop (Fig. 6), these errors need to

be detected. Depending on the application, the controller may or may not be able to deal with erroneous or missing values. Techniques to estimate missing values that have been reported in the literature include reference days (Cools et al. 2007), multiple imputation (Little and Rubin 1987), time series analysis (Cools et al. 2007), and Kalman and particle filtering (Hegyi et al. 2006; Wang and Papageorgiou 2005).

For the purpose of freeway traffic control, the most commonly measured traffic variables are the traffic flow, the speed, and the occupancy of the freeway traffic. The choice of these variables is influenced by their importance in traffic theory as well as by the ease by which they can be measured with most common traffic detector technologies.

There exists a wide variety of technologies (Kell et al. 1990) to measure traffic variables such as pneumatic sensors, inductive loops, cameras, ultrasonic sensors, microwave sensors, active and passive infrared sensors, passive acoustic arrays, and magnetometers.

Inductive loops are the most widespread detection systems to date and were introduced as traffic detection systems in the 1960s (Middleton et al. 2002). The main advantages of inductive loops are their wide application range, the flexible design, and the availability of the common traffic variables. The main disadvantages of inductive loop technology are the sensitivity to wear and tear due to physical stress on the loops induced by traffic, the susceptibility of the loops to damage by road maintenance works, and the special installation and maintenance requirements (such as lane closure during maintenance) (Kell et al. 1990).

Traffic detection using video cameras emerged during the 1980s and is a nonintrusive technology that is becoming more and more popular (Middleton et al. 2002). A fixed camera is mounted above the freeway and its images are sent to a video processing unit that extracts the desired variables using image processing algorithms. Camera traffic detection technology can provide the common traffic variables, is less prone to wear and tear by traffic, and generally requires less lane closures for maintenance and reconfiguration. The main disadvantages of traffic cameras are the higher upfront cost of the installation compared to inductive loops and the

dependence on visibility conditions (such as fog, heavy snow, sunlight shining directly into the camera could heavily impair the quality of the images taken by the traffic cameras). Other factors that adversely affect the detection accuracy are vibrations caused by wind and traffic, lack of contrast between vehicle and road color, and varying lighting conditions (such as during dusk and dawn) (Federal Highway Administration, US Department of Transportation 1995b; Klein et al. 1997, 2006).

In addition to the registration of the traditional traffic variables, video camera technology is also applied to register travel times on corridors or to obtain information regarding the routes followed throughout the network. This information can be obtained by tracking the vehicles at strategic locations (such as at large junctions or at the entrances and the exits of the area under consideration) using automated license plate recognition algorithms, for example. These systems consist of video image processing units connected to video cameras monitoring the traffic. Often, these systems are implemented at sites where the hardware to register the vehicles is already available (e.g., automated toll booths) (Castello et al. 1999; Soriguera et al. 2007).

Estimation

In order to coordinate or to integrate traffic control measures, the spatial aspect of the traffic network needs to be taken into account such that the impact of control measures on distant parts of the traffic network can be accounted for. However, the traffic sensors discussed above are traffic sensors that are localized in space, and as a consequence, they only yield information on the evolution of the traffic state on the freeway through time and at the sensor locations. Implementing traffic detectors very densely on the freeway in order to register the traffic states for every freeway section would be inconvenient and costly. However, data fusion techniques (such as extended Kalman filtering or particle filtering) allow one to combine traffic measurements scattered over the traffic network in order to obtain network traffic state estimation (Hegyi et al. 2006; Wang and Papageorgiou 2005; Wang et al. 2006). Traffic state estimation and prediction can be used in the

implementation of control measures as is illustrated by simulation in Bellemans et al. (2006b).

The similarities between traffic flows and flows of compressible fluids have since long been documented in the literature (Lighthill and Whitham 1955; Richards 1956) (see also section “Traffic Flow Modeling” below). However, when it comes to the routing of traffic flows in networks, a major difference emerges; while the particles in a fluid have no predetermined destination, the vehicles traveling through a traffic network are traveling from a particular origin to a particular destination. Hence, the destination of the vehicles constrains the alternative routes that can be chosen. It is clear that route guidance control measures have an impact on the routing process. However, since travel times are typically an important factor in the routing process of well-informed travelers, other traffic measures, in combination with the traffic demands, influence the routing behavior as well. In order to assess the impact of traffic measures on the traffic states in the traffic network, the traffic demands (OD matrices) and the routing process need to be modeled. As the traffic demand cannot be measured directly, it needs to be estimated using the traffic measurements in the traffic network. Several techniques to estimate the traffic demands (OD matrices) that correspond to the measured traffic states in the traffic network have been developed. For an overview of the literature, the interested reader is referred to Lin and Chang (2007). The route choice process, which influences the impact of control measures through rerouting effects, is also the subject of research (Karimi et al. 2004).

Traffic Flow Modeling

Traffic flow models can be classified according to various criteria such as *area of application*, *level of detail*, and *deterministic versus stochastic* (Hoogendoorn and Bovy 2001).

An example of the application of traffic flow models for the design of traffic control measures is model predictive control, which makes use of an internal prediction model in order to find the best traffic control measures to be applied to the real

traffic process. Since these models are operated in real-time, and are often used to evaluate several control scenarios, they should allow for fast execution on a computer.

For the assessment of traffic control strategies often a simulation model is used instead of (or before) a real-world test. Simulation has several advantages. Above all, simulation is cheaper and faster, and it does not require real human drivers as test subjects. It also provides an environment where the unpredictable disturbances of a field test, such as weather influences, traffic demand variations, and incidents, can be excluded, or if necessary simulations can be repeated under exactly the same disturbance scenario.

Since none of the available traffic models perfectly describes the real traffic behavior, one has to keep in mind the intended application, when making the choice between the available traffic flow models. As Papageorgiou (1998) argues for macroscopic traffic flow models an important criterion is that the model should have sufficient descriptive power to reproduce all important phenomena for the intended application. Similar arguments are also used by Kerner (2004) but for different phenomena.

Traffic models can also be classified according to the level of detail with which they describe the traffic process:

- **Microscopic** models describe the behavior of individual vehicles. Important aspects of microscopic models are the so-called *car following* and *lane changing* behavior. Car following and lane changing behavior is generally described as a function of the distance to and (relative) speed of the surrounding vehicles, and the desired speed. Since the vehicles are modeled individually in microscopic traffic models, it is easy to assign different characteristics to each vehicle. These characteristics can be related to the driving style of the driver (aggressive, patient), vehicle type (car, truck), its destination, and route choice.

A special type of microscopic traffic models are the *cellular-automaton* models (Kerner 2004; Maerivoet and De Moor 2005; Wolf

1999) in which the freeway is discretized into cells of about 7.5 m length. Each cell can contain only one vehicle and the traffic dynamics is described by a probabilistic model of the hopping behavior of the vehicles through the cells.

In general, it is difficult to calibrate microscopic models with real traffic data, due to the large number of parameters in this type of models and the poor availability of appropriate traffic data.

We refer the interested reader to Algers et al. (2000) for an extensive comparison of commercial microscopic simulation models and to Hoogendoorn and Bovy (2001) for a more theoretical overview.

- **Mesoscopic** models do not track individual vehicles, but describe the behavior of individual vehicles in probabilistic terms. Examples of mesoscopic models are headway distribution models (Branston 1976) and gas-kinetic models (Hoogendoorn and Bovy 2000). Typically, these models are not used for traffic control.
- **Macroscopic** models use a high level of aggregation without distinguishing between individual vehicle behaviors. Instead, traffic is described in aggregate terms as average speed, average flow, and average density.

Macroscopic traffic flow modeling started when Lighthill and Whitham (1955) presented in 1955 a model based on the analogy between traffic flows and flows in rivers. Independently of Lighthill and Whitham 1 year later Richards (1956) published a similar model. Therefore, this model is usually referred to as the Lighthill-Whitham-Richards (LWR) model.

Since 1955 a large variety of macroscopic traffic flow models have evolved from the LWR model with differences in the order of the model, the phenomena that they (re)produce (such as capacity drop, stop-and-go waves, and other congestion phenomena or patterns), and the effects of heterogeneous traffic (cars, trucks, etc.) (Daganzo 1994; Helbing 1997; Hoogendoorn and Bovy 2000; Payne 1971).

Another approach has been followed by Kerner (2004) who developed a qualitative traffic flow theory (and a related microscopic traffic flow model) based on empirical observation. This theory distinguishes three so-called traffic phases: free-flow, synchronized flow, and jammed traffic, and describes the transition between these phases qualitatively in probabilistic terms.

A last classification that is relevant in the context of traffic control is whether the model is deterministic or stochastic. Deterministic models define a relationship between model inputs, variables, and outputs that typically describe the average behavior of traffic. Stochastic models describe traffic behavior in terms of relationships between random variables, e.g., random reaction time of drivers, randomness in equilibrium speed-density (or car following) relationships, and route choice. These stochastic effects can reproduce phenomena such as the creation of traffic jams by random fluctuations in traffic flows (Smulders 1990) and can be used for the stochastic evaluation of traffic control approaches. Another application of stochastic traffic flow models is in the area of state estimation, which is an essential part of control approaches such as optimal control or model predictive control (Hegyi et al. 2006; Wang and Papageorgiou 2005).

Freeway Traffic Control Measures

In this section we give an overview of control measures that are used or could be used to improve traffic performance. We focus on control measures that are currently applied, or could be applied in the near future, such as ramp metering, speed limits, and route guidance. For each control measure we present the principle of operation including the control approaches, and the existing field tests and simulation results. At the end of this section some other traffic control measures are presented that may also be used to improve the performance of traffic systems.

Ramp Metering

Principle of Operation

Ramp metering is one of the most investigated and applied freeway traffic control measures. A ramp metering setup is implemented as a traffic signal that is placed at the on-ramp of a freeway as shown in Fig. 7. The required metering rate is implemented by appropriately choosing the phase lengths of the traffic signal. Several ramp metering implementations can be distinguished (Chaudhary and Messer 2000), e.g., single-lane with one vehicle per green ramp metering, single-lane with multiple vehicles per green ramp metering (bulk metering), and dual-lane ramp metering.

Ramp metering can be used in two modes: the *traffic spreading mode* and the *traffic restricting mode*. In the traffic spreading mode ramp metering smooths the merging process of on-ramp traffic by breaking the platoons and by spreading the on-ramp traffic demand over time as observed by Elefteriadou (1997). This mitigates the shock waves that can occur under high traffic density conditions. In this application the metering rate equals the average arrival rate of the vehicles.

Restrictive ramp metering can be used for three different purposes:



Freeway Traffic Management and Control, Fig. 7 Ramp metering at the freeway A13 in Delft, The Netherlands. One car may pass per green phase. To prevent red-light running the control is enforced

- **Prevention of breakdowns.** When traffic is dense, ramp metering can prevent a traffic breakdown on the freeway by adjusting the metering rate such that the density on the freeway remains below the critical value. Preventing a traffic breakdown has not only the advantage that it results in a higher flow but also that it prevents the creation of a jam that could block the off-ramp upstream the on-ramp (as shown in Fig. 4). These effects are studied in detail by Papageorgiou and Kotsialos (2002). Daganzo (1996) has quantified the role of ramp metering in avoiding the activation of freeway gridlocks.
- **Influencing route choice.** Ramp metering can be implemented to influence the traffic demand and traffic routing. The impact of ramp metering on the traffic state and on the travel times is taken into account by the drivers in their routing behavior (Zhang 2007). Banks (2005) has described a theory to apply ramp metering to influence traffic routing to avoid freeway bottlenecks. Based on a similar idea Middelham (1999) has performed a synthetic study on the route choice effects of ramp metering.
- **Localization of traffic jams.** According to Kerner (2004) ramp metering can prevent the backpropagation of traffic jams and shock waves occurring at on-ramps. This could be beneficial on the network level since it could localize the traffic jam, and it could also be beneficial to the traffic throughput.

The control strategies that have been developed for restrictive ramp metering can be classified as static or dynamic, fixed-time or traffic-responsive, and local or coordinated.

Fixed-time strategies use (time-dependent) fixed metering rates that are determined off-line based on historical demands. This approach was first suggested by Wattleworth (1965) and was extended to a dynamic traffic model by Papageorgiou (1980). The disadvantage of fixed-time strategies is that they do not take into account the day-to-day variations in the traffic demand or the variations in the demand during a period with a constant

metering rate, which may result in underutilization of the freeway or inability to prevent congestion.

Traffic-responsive strategies solve these issues by adjusting on-line the metering rate as a function of the prevailing traffic conditions. These strategies also aim at the same objectives as the fixed-time strategies but use direct traffic measurements instead of historical data to prevent or to reduce congestion. One of the best known strategies is the *demand-capacity* strategy (Papageorgiou and Kotsialos 2000):

$$q_{\text{ramp}}(k) = \begin{cases} q_{\text{cap}} - q_{\text{in}}(k-1) & \text{if } o_{\text{meas}}(k-1) \leq o_{\text{cr}} \\ q_{\text{r}, \text{min}} & \text{otherwise} \end{cases}$$

with $q_{\text{ramp}}(k)$ the admitted ramp flow at time step k , q_{cap} the downstream freeway capacity, $q_{\text{in}}(k)$ the freeway flow measured upstream of the on-ramp at time step k , $q_{\text{r}, \text{min}}$ the minimal on-ramp flow during congestion, $o_{\text{meas}}(k)$ the occupancy downstream the on-ramp at time step k , and o_{cr} is the critical occupancy (at which the flow is maximal). Since the traffic state on the freeway cannot be determined based on the measurement of the traffic flow alone, the downstream occupancy is measured in order to determine whether congestion is present ($o_{\text{meas}}(k-1) > o_{\text{cr}}$) or not.

A similar strategy is occupancy-based ramp metering, where the upstream traffic flow measurement from demand-capacity ramp metering is replaced by an occupancy measurement. The measured occupancies are related to traffic flows based on historical measurements. Next, the demand-capacity approach described above can be applied (Carvell et al. 1997). A common disadvantage of both demand-capacity formulations is that they have an (open-loop) feedforward structure, which is known to perform poorly under unknown disturbances and cannot guarantee a zero offset in the output under steady-state conditions.

A better approach is to use a (closed-loop) feedback structure, because it allows for controller formulations that can reject disturbances and have zero steady-state error. ALINEA (Papageorgiou

et al. 1991) is such a ramp metering strategy and its control law is defined as follows:

$$q_{\text{ramp}}(k) = q_{\text{ramp}}(k-1) + K(\hat{o} - o_{\text{meas}}(k)),$$

where $q_{\text{ramp}}(k)$ is the metered on-ramp flow at time step k , K is a positive constant, $\hat{o}$ is a set-point for the occupancy, and $o_{\text{meas}}(k)$ is the measured occupancy on the freeway downstream of the on-ramp at time step k . ALINEA tries to maintain the occupancy on the freeway equal to a set-point $\hat{o}$, which is chosen in the region of stable operation. Given the probabilistic nature of traffic operation, the set-point $\hat{o}$ is often chosen somewhat smaller than the critical occupancy in order to guarantee free-flow traffic operation.

More advanced ramp metering strategies are the traffic-responsive coordinated strategies such as METALINE (Papageorgiou et al. 1990b), FLOW (Jacobson et al. 1989), or methods that use optimal control (Kotsialos et al. 2001) or model predictive control (Belleman 2003).

The ramp metering strategies discussed above attempt to conserve free-flow traffic on the freeway. However, given the probabilistic nature of traffic operation, congestion can set in at lower or higher densities than the critical density. Based on these insights, Kerner (2004, 2007a) defined a *congested-pattern control approach* to ramp metering called ANCONA. The basic idea of ANCONA is to allow congestion to set in, but to keep congested traffic conditions to the minimum level possible. Once congestion sets in, ANCONA tries to reestablish free-flow conditions on the freeway by reducing the on-ramp metering rate. Kerner claims that, by allowing congestion to set in, ANCONA utilizes the available freeway capacity better. The control rule of ANCONA is given by Kerner (2004):

$$q_{\text{ramp}}(k) = \begin{cases} q_1 & \text{if } v_{\text{det}}(k) \leq v_{\text{cong}} \\ q_2 & \text{if } v_{\text{det}}(k) > v_{\text{cong}} \end{cases},$$

where $q_{\text{ramp}}(k)$ is the on-ramp flow at time step k , q_1 and q_2 are heuristically determined constant flows with $q_1 < q_2$, $v_{\text{det}}(k)$ is the traffic speed on the freeway just upstream of the on-ramp at time

step k , and v_{cong} is the speed threshold that separates the free and the synchronized (locally congested) traffic flow phases.

Field Tests and Simulation Studies

Several field tests and simulation studies have shown the effectiveness of ramp metering. In Paris on the Boulevard Périphérique and in Amsterdam several ramp metering strategies have been tested (Papageorgiou et al. 1997). The demand-capacity, occupancy, and ALINEA strategies were applied in the field tests at a single ramp in Paris. It was found that ALINEA was clearly superior to the other two in all the performance measures (total time spent, total traveled distance, mean speed, mean congestion duration). At the Boulevard Périphérique in Paris the multivariable (coordinated) feedback strategy METALINE was also applied and was compared with the local feedback strategy ALINEA. Both strategies resulted in approximately the same performance improvement (Papageorgiou et al. 1990b). One of the largest field tests was conducted in the Twin Cities metropolitan area of Minnesota. In this area, 430 operational ramp meters were shut down to evaluate their effectiveness. The results of this test show that ramp metering not only serves the purposes of improving traffic flow and traffic safety but also improves travel time reliability (Cambridge Systematics, Inc. 2001; Levinson and Zhang 2006).

A number of studies have simulated ramp metering for different transportation networks and traffic scenarios, with different control approaches, and with the use of microscopic and macroscopic traffic flow models (Hasan et al. 2002; Hellinga and Van Aerde 1995; Kotsialos et al. 2001; Papageorgiou et al. 1990a, 1991; Taylor et al. 1998). Generally the total network travel time is considered as the performance measure and is improved by about 0.39–30% when using ramp metering.

The validation of ANCONA versus ALINEA is performed by simulation by Kerner (2007a), who found that ANCONA in some cases can lead to higher flows.

Also note that Kerner has criticized modeling approaches to simulations of freeway traffic control strategies in Kerner (2007b) (see also Papageorgiou et al. (2007) for some comments). The fundamental difference between Kerner's view and many classical approaches is that Kerner concluded, based on empirical observations, that the capacity of a bottleneck should be described as a range of capacities, between a minimum value $C_{\min}$ and a maximum value $C_{\max}$, and that all capacities in this range are considered to hold *at the same time*, and thus any flow value in this range might cause a breakdown. In Kerner's theory a breakdown in free flow is caused by the appearance of a so-called nucleus, a local disturbance that is large enough to cause a breakdown. Since the classical approaches don't consider nucleus formation, they are, according to Kerner, invalid. We refer the interested reader to Kerner (2009, 2013, 2015, 2016) and Kerner et al. (2015) for further details on Kerner's criticism on the classical approaches.

Dynamic Speed Limits

Dynamic speed limits are used to reduce the maximum speed on freeways according to given performance, safety, or environmental criteria. An example of a dynamic speed limit gantry is shown in Fig. 8.



Freeway Traffic Management and Control, Fig. 8 A variable speed limit gantry on the A13 freeway near Overschie, The Netherlands. In this particular case the maximum speed limit is 80 km/h due to environmental reasons, and the limit may drop to 70 km/h or 50 km/h in case of a downstream jam

Principle of Operation

The working principle of speed limit systems can be categorized based on their intended effects: improving safety, improving traffic flow, or their environmental effects, such as reducing noise or air pollution.

It is generally accepted that speed reduction on freeways leads to improved safety (Sisiopiku 2001; Transportation Research Board 1998; Wilkie 1997). Lower speeds in general are associated with lower crash rates and with a lower impact in case of a collision. If the environmental conditions or traffic conditions are such that the posted maximum speeds are considered to be unsafe, the speed limit can be lowered to match the given conditions. Dynamic speed limits may function as a warning that an incident or jam is present ahead.

In the literature, basically two approaches to dynamic speed limit control can be found for flow improvement. The first emphasizes the homogenization effect (Alessandri et al. 1999; Hardman 1996; Kühne 1991; Smulders 1990; van den Hoogen and Smulders 1994), whereas the second is more focused on preventing traffic breakdown by reducing the flow by means of speed limits (Chien et al. 1997; Hegyi et al. 2005b; Lenz et al. 2001).

- The basic idea of homogenization is that speed limits can reduce the speed (and/or density) differences, by which a more stable (and safer) flow can be achieved. The homogenizing approach typically uses speed limits that are above the critical speed (i.e., the speed that corresponds to the maximal flow). So, these speed limits do not limit the traffic flow, but only slightly reduce the average speed (and slightly increase the density). In theory this approach can increase the time to breakdown slightly (Smulders 1990), but it cannot suppress or resolve shock waves. An extended overview of speed limit systems that aim at reducing speed differentials is given by Wilkie (1997).
- The traffic breakdown prevention approach focuses more on preventing too high densities, and also allows speed limits that are lower than

the critical speed in order to limit the inflow to these areas. By resolving the high-density areas (bottlenecks) higher flow can be achieved in contrast to the homogenization approach (Hegyi et al. 2005b).

Currently, the main purpose of most of the existing practical dynamic speed limit systems is to increase safety by lowering the speed limits in potentially dangerous situations, such as upstream of congested areas or during adverse weather conditions (Sisiopiku 2001; Transportation Research Board 1998; Wilkie 1997). Although these systems primarily aim at safety, in general they also have a positive effect on the flow, due to the fact that preventing accidents results in a higher flow. There are also some examples of practical systems that are designed with the purpose of flow improvement (Rees et al. 2004; Schik 2003) – with varying success. These practical systems in general use a switching scheme based on traffic conditions, weather conditions, visibility conditions, or pavement conditions (Rees 1995; Williams 1996).

Several control methodologies are used in the literature to find a control law for speed control, such as multilayer control (Li et al. 1995), sliding-mode control (Lenz et al. 2001), and optimal control (Alessandri et al. 1999). In Di Febbraro et al. (2001), optimal control is approximated by a neural network in a rolling horizon framework. Other authors use, or simplify their control law to, a control logic where the switching between the speed limit values is based on traffic volume, speed, or density (Hardman 1996; Kühne 1991; Lenz et al. 2001; Smulders 1990; van den Hoogen and Smulders 1994). We refer the interested reader for further reading about the various control methodologies to the references at the end of this entry.

Some authors recognize the importance of anticipation in the speed control scheme. A pseudo-anticipative scheme is used in Lenz et al. (2001) by switching between speed limits based on the density of the neighboring downstream segment. Explicit predictions are used in Alessandri et al. (1999) and Di Febbraro et al. (2001), and this is the only approach that results in a significant flow improvement. The heuristic

algorithm proposed in Wilkie (1997) also contains anticipation to shock waves being formed.

Another concept of dynamic speed limits is their use in combination with ramp metering to prevent a breakdown on the freeway at the on-ramp and to prevent the ramp queue to propagate back to the urban network (Hegyi et al. 2005a) by taking over the flow limitation function from the ramp metering when the ramp queue has reached its limit.

Field Tests and Simulation Studies

Field data evaluations show that in general homogenization results in a more stable and safer traffic flow, but no significant improvement of traffic volume is expected nor measured (Schik 2003; van den Hoogen and Smulders 1994). Since the introduction of speed control on the M25 in the United Kingdom, an increase of flow of 1.5% per year is reported for the morning peaks, but no improvement is found in the afternoon peaks (Rees et al. 2004).

The effect of dynamic speed limits on traffic behavior strongly depends on whether the speed limits are enforced or not, and on whether the speed limits are advisory or mandatory, which also determines the suitability for a certain application. Most application oriented studies (Smulders and Helleman 1998; van den Hoogen and Smulders 1994; Wilkie 1997) enforce speed limits, except for Kühne (1991). Enforcement is usually accepted by the drivers if the speed limit system leads to a more stable traffic flow.

Route Guidance

Principle of Operation

Route guidance systems assist drivers in choosing their route when alternative routes exist to their destination. The systems typically display traffic information such as congestion length, the delay on the alternative routes, or the travel time to the next common point on the alternative routes (an example is given in Fig. 9). Recently, in-car navigation system manufacturers have shown interest in providing route advice taking the traffic jams and travel times on the alternative routes into account.



Freeway Traffic Management and Control, Fig. 9 A route guidance system in The Netherlands showing traffic jams lengths on alternative routes to Schiphol Airport (Photo courtesy of Peek Traffic B.V.)

In route guidance the notions of *system optimum* and *user equilibrium* (or *user optimum*) play an important role. The system optimum is achieved when the vehicles are guided such that the total costs of all drivers (typically the total travel time) is minimized. However, the system optimum does not necessarily minimize the travel time (or some generalized cost measure) for each individual driver. So, some drivers may select another route that has a shorter individual travel time (lower cost). The traffic network is in user equilibrium when on each utilized route the costs are equal, and on routes that are not utilized the cost is higher than that on the utilized routes. This means that no driver has the possibility to find another route that reduces his or her individual cost.

The cost function is typically defined as the travel time, either as the *predicted travel time* or as the *instantaneous travel time*. The predicted travel time is the time that the driver will experience when he or she drives along the given route, while the instantaneous travel time is the travel time determined based on the current speeds on the route. In a dynamic setting the speeds in the network may change during a trip, and consequently the instantaneous travel time may be different from the predicted travel time.

Papageorgiou and Messmer (1991) have developed a theoretical framework for route guidance in traffic networks. Three different traffic

control problems are formulated: an optimal control problem to achieve the system optimum (minimize the total time that is spent in the network), an optimal control problem to achieve a user optimum (equalize travel times), and a feedback control problem to achieve a user optimum (equalize travel times). The resulting control strategies are demonstrated on a test network with six pairs of alternative routes. The feedback control strategy is tested with instantaneous travel times and results in a user equilibrium for most alternative routes, and the resulting total time spent by the vehicles in the network is very close to the system optimum.

Wang et al. (2003) combine the advantages of a feedback approach (relatively simple, robust, fast) and predicted travel times. The resulting *predictive feedback* controller is compared with optimal control and with a feedback controller based on instantaneous travel times. When the disturbances are known the simulation results show that the predictive feedback results in nearly optimal splitting rates, and is clearly superior to the feedback based on instantaneous travel times. The robustness of the feedback approach is shown for several cases: incorrectly predicted demand, an (unpredictable) incident, and an incorrect compliance rate.

Field Tests and Simulation Studies

In several studies, it is assumed that the turning rates can be directly manipulated by route guidance messages (Papageorgiou and Messmer 1991; Wang et al. 2003). In the case of in-car systems, it is plausible that by giving direct route advice to individual drivers the splitting rates can be influenced sufficiently. However, in the case of route guidance by variable message signs the displayed messages do not directly determine the splitting rates: The drivers make their own decisions. Therefore, empirical studies about the reaction of drivers to dynamic route information messages and the effectiveness of route guidance can provide useful information.

Kraan et al. (1999) present an extensive evaluation of the impact on network performance of variable message signs on the freeway network around Amsterdam, The Netherlands. Several performance indicators are compared before and

after the installation of 14 new variable message signs (of which seven are used as incident management signs and seven as dynamic route information signs). The performance indicators, such as the total traveled distance, the total congestion length and duration, and the instantaneous travel time delay are compared for alternative routes and for most locations a small but statistically significant improvement is found. The day-to-day standard deviation of these indicators decreased after the installation of the variable message signs, which indicates that the travel times have become more reliable.

Another field test is reported by Diakaki et al. (2000) in which a combination of route guidance, ramp metering, and urban traffic control is applied to the M8 corridor network in Glasgow, UK. The applied control methodology resulted in an increased network throughput and in a reduced average travel time.

Other Control Measures

Besides ramp metering, dynamic speed limits, and route guidance, there are also other dynamic traffic control measures that can potentially improve the traffic performance. In this section we describe a selection of such measures and describe in which situations they are useful (cf. Middelham 2003).

- **Peak lanes.** During peak hours the hard shoulder lane of a freeway (which is normally used only by vehicles in emergency) is opened for traffic. Whether the lane is opened or closed is communicated by variable message signs showing a green arrow or a red cross. Due to the extra lane the capacity of the road is increased, which could prevent congestion. The disadvantage of using the emergency lane as a normal lane is that the safety may be reduced. For this reason, often extra conditions ensuring safety are required, such as the creation of emergency refuges adjacent to the hard shoulder lane, or the requirement that emergency services should be able to access the incident location over or through the guard rail. Furthermore, there may be CCTV surveillance or vehicle patrols to detect incidents early. This traffic control measure is useful

where the additional capacity prevents congestion and the downstream infrastructure can accommodate the increased traffic flow.

- **Dedicated lanes.** During congestion the shoulder lane may be opened for dedicated vehicles, such as public transport, freight transport, or high occupancy vehicles (with more than two passengers). This reduces the hindrance that congestion causes to these vehicles. Furthermore, public transport can be made more reliable and thus more attractive by this measure. A dedicated freight transport lane increases the stability and homogeneity of the traffic flow.
- **Tidal flow.** Tidal flow allows one to use a freeway lane in the one or the other direction. Depending on the direction of the highest traffic demand the direction of operation is determined. This direction is communicated by a variable message sign showing a red cross or a green arrow. This traffic control measure is useful when the traffic demand is typically not high in both directions simultaneously.
- **The “keep your lane” directive.** When the “keep your lane” directive is displayed, the drivers are not allowed (not recommended) to change lanes. This results in less disturbances in the freeway traffic flow, which may prevent congestion. This traffic control measure is useful when the traffic flow is nearly unstable (close to the critical density) and may be a good alternative to homogenizing speed limits.

Network-Oriented Traffic Control Systems

The integration of traffic control measures in freeway networks is essential in order to ensure that the control actions taken at different locations in the network reinforce rather than counteract or even cancel each other. While in the previous section individual traffic control measures were discussed along with the most prevalent local control strategies, this section explicitly considers the integration of several control measures in a freeway network context.

Although the focus in this entry is on automatic control systems, it must be noted that in practice in

traffic control centers there is also often a human controller with “oversight” of the system as a safeguard against problems with the system and in view of the complexity of the control problem.

In network-oriented traffic control two ingredients play an important role: *coordination* and *prediction*. Since in a dense network the effect of a local control measure could also influence the traffic flows in more distant parts of the network, the control measures should be coordinated such that they serve the same objectives. Taking into account the effects of control measures on distant parts of the network often also involves prediction, due to the fact that the effect of a control measure has a delay that is at least the travel time between the two control measures in the downstream direction, and at least the propagation time of shock waves in the upstream direction. An advantage of control systems that use explicit predictions is that by anticipating on predictable future events the control system can also *prevent* problems instead of only *reacting* to them. However, it must be noted that while all network-oriented control approaches apply some form of coordination, many approaches do not explicitly make use of predictions.

Network-oriented traffic control has several advantages compared to local control since it ensures that local traffic problems are solved with the aim of achieving an improvement on the network level. For example, solving a local traffic jam only can have as consequence that the vehicles run faster into another (downstream) jam, whereas still the same amount of vehicles have to pass the downstream bottleneck (with a given capacity). In such a case, the average travel time on the network level will still be the same, regardless of whether or not the jam is solved. However, a global approach would take into account both jams and, if possible, solve both of them.

Furthermore, network-oriented control approaches can utilize network-related historical information. For example, if dynamic OD data is available, control on the network level can take advantage of the predictions of the flows in the network. Local controllers are not able to optimize the network performance even if the dynamic OD data is available, because the effect of the control

actions on downstream areas is not taken into account.

In the literature basically three approaches exist for coordinating traffic control measures: model-based optimal control methods, knowledge-based methods, and methods that use simple feedback or switching logic, for which the parameters are optimized. In some approaches different methods are combined in a hierarchical control structure. We discuss these approaches in the following subsections.

Model-Based Control Methods

Model-based traffic control techniques use a traffic flow model for predicting the future behavior of the traffic system based on

- The current state of traffic
- The expected traffic demand on the network level, possibly including OD relationships and external influences, such as weather conditions
- The planned traffic control measures

Since the first two items cannot be influenced (except for the possibility that based on real-time congestion information people cancel their planned trip, change the departure time, or travel via another modality), the future performance of the traffic system is optimized by selecting an appropriate scenario for the traffic control measures. Methods that use *optimal control* or *model predictive control* explicitly take the complex nonlinear nature of traffic into account. For example, they take into account the fact that the effect of ramp metering on distant on-ramps will be delayed by the (time-varying) travel time between the two on-ramps. In general, the other existing methods (such as knowledge-based methods, or control parameter optimization) do not explicitly take this kind of delay into account. Furthermore, other advantages of the model-based methods are that traffic demand predictions can be utilized, constraints on the ramp metering rate and the ramp queues can be included easily, and a user-supplied objective function can be optimized.

Kerner has developed a network traffic routing approach that builds upon his view of stochastic

breakdown at freeway bottlenecks. His approach aims to maximize the probability of not having any breakdown in a network with multiple bottlenecks (Kerner 2017).

Optimal control has been successfully applied in simulation studies to integrated control of ramp metering and freeway-to-freeway control (Kotsialos et al. 2001), to route guidance (Hoogendoorn 1997), and to integration of ramp metering and route guidance (Kotsialos et al. 2001, 2002). In Kotsialos et al. (2001, 2002) the integrated controller performed better than route guidance or ramp metering alone.

The model predictive control (MPC) approach is an extension of the optimal control, which uses a rolling horizon framework. This results in a closed-loop (feedback) controller, which has the advantage that it can handle demand prediction errors and disturbances (such as incidents). MPC is computationally more efficient than optimal control due to the shorter prediction and control horizons, and it can be made adaptive by updating the prediction model on-line.

MPC-based control has been applied in simulations to coordinated ramp metering (Bellemans et al. 2006a), to integrated control of ramp metering and dynamic speed limits (Hegyi et al. 2005a), and to integrated control of ramp metering and route guidance (Karimi et al. 2004; van den Berg et al. 2005). An illustration of the ability of MPC-based traffic control to deal with a model mismatch was given in Bellemans et al. (2006b). In Bellemans (2003), it was illustrated by simulation of a simple proof-of-concept network that MPC can be implemented to account for the rerouting behavior of vehicles due to changing travel times caused by applying ramp metering.

Knowledge-Based Methods

Knowledge-based traffic control methods typically describe the knowledge about the traffic system in combination with the control system in terms that are comprehensible for humans. Given the current traffic situation the knowledge-based system generates a solution (control measure) via reasoning mechanisms. A typical motivation for these systems is to help traffic control center operators to find good (not necessarily the best)

combinations of control measures. The operators often suffer from cognitive overload by the large number of possible actions (control measures) or by time pressure in case of incidents. The possibility for the operators to track the reasoning path of the knowledge-based system makes these systems attractive and more convincing.

An example of a knowledge-based system is the TRYS system (Cuenca et al. 1995; Hernández et al. 1999; Molina et al. 1998), which uses knowledge about the physical structure of the network, the typical traffic problems, and about effects of the available control measures. The TRYS system has been installed in traffic control centers in Madrid and Barcelona, Spain.

Another knowledge-based system is the freeway incident management system (Gupta et al. 1992) developed in Massachusetts, which assists in the management of nonrecurrent congestion. The system contains a knowledge base and a reasoning mechanism to guide the traffic operators through the appropriate questions to manage incidents. Besides incident detection and verification, the system assists in notifying the necessary agencies (such as ambulance, clean-up forces, towing company) and in applying the appropriate diversion measures. The potential benefits (reduced travel times by appropriate diversion) are illustrated by a case study on the Massachusetts Turnpike. The knowledge-based expert system called freeway real-time expert-system demonstration (Ritchie 1990; Zhang and Ritchie 1994) has similar functionalities and is illustrated by applying it to a section of the Riverside Freeway (SR-91) in Orange County, California.

Control Parameter Optimization Methods

Alessandri et al. (1999) follow another approach: A relatively simple control law is used for speed limit control and ramp metering, and the parameters of the control law are found by simulating a large number of scenarios and optimizing the average performance. In Alessandri et al. (1999) a dynamic speed limit switching scheme is developed. The speed limits switch between approximately 70 and 90 km/h, and the switching is based on the density of the segment to be controlled and two thresholds (to switch up and to switch down).

The switching scheme uses a hysteresis loop to prevent too frequent switching. Optimizing the thresholds for several objectives resulted in a slight increase of the average throughput, a decrease of the sum of squared densities – which can be considered as a measure of inhomogeneity (since a nonuniform distribution of vehicles over a freeway stretch results in a higher sum of squared densities) – and a small decrease of the total time spent by the vehicles in the network.

Hierarchical Control

The increasing number of traffic control measures that need to be controlled in a network-control context, as well as their interactions, drastically increases the computational complexity of computing the optimal control signals. Hierarchical control was introduced by some authors in order to tackle this problem (Chen et al. 1997; Kotsialos and Papageorgiou 2005; Papageorgiou 1983). In hierarchical control the controlled process is partitioned in several subprocesses, and the control task is performed by a high-level controller and several low-level controllers. The high-level controller determines centrally the set-points or trajectories representing the desired behavior of the subprocesses. The low-level controllers are used to steer the subprocesses according to the set-points or trajectories supplied by the high-level controller. Usually, the high-level controller operates at a slower timescale than the low-level controllers. Hierarchical systems do not only enable coordination of control for large networks, but they also provide high reliability and robustness (Hotz et al. 1992).

Future Directions

Although there is a large interest in developing freeway traffic control systems, there is by no means a consensus about the most suitable approaches or methods. One of the reasons is that traffic phenomena, such as traffic breakdown and jam resolution, are not perfectly understood (Kerner 2004) and different views lead to different approaches. In addition, technological developments such as advanced sensor technologies and

intelligent vehicles open new possibilities that enable or require new control approaches.

Advanced Sensor Technologies

Given the complexity of traffic state estimation, traffic demand estimation, and the collection of routing information based on conventional traffic measurement data, new data collection methods are being investigated.

Instead of registering vehicles at certain locations using hardware on the freeway network, floating car data can be collected. The collection of floating car data, where individual vehicles are tracked during their journey through the network, provides valuable route choice and traffic demand information. The evolution in mobile computing and in mobile communication has enabled the incorporation of these technologies in the field of traffic data collection, allowing more detailed and more cost-effective data collection. In contrast to the traditional data collection methods that were discussed in section “[Measurements](#),” this section deals with two data collection methodologies that are enabled by mobile computing and communication.

Cell phone service providers collect data regarding the base station each cell phone connects to and the time instant the connection is initiated. Since cell phones regularly connect to their current base station and since the location of these base stations is known, information about the journey of the cell phone can be extracted from the service provider’s database. By monitoring a large number of cell phones, and more in particular their hands-off processes when hopping from one base station to the next, an impression of the traffic speeds and the travel times can be obtained (Smith et al. 2004).

The global positioning system (GPS) is well-suited for tracking probe vehicles through space and time in order to obtain route information and travel times (Byon et al. 2006; Taylor et al. 2000). With the further miniaturization of electronics, the processing power available in mainstream navigation units and mobile data communication facilities (e.g., GPRS) the cost of instrumenting fleets of probe vehicles decreases. For example, fleets of taxis, buses, and trucks can be used as probe

vehicles as they are often readily equipped with GPS and data communication technology. When dealing with probe vehicles, care must be taken to ensure that the number of probe vehicles is large enough in order to be able to accurately determine the traffic state (Jiang et al. 2006).

Although the technologies presented above are readily available and have been used in the past, their structural deployment as a source for traffic measurements for dynamic traffic control systems still needs to break through. Some issues that may determine whether floating car data becomes a viable option for large-scale data collection are the accuracy of the data obtained, privacy concerns related to registering the whereabouts of individuals, operational communication and computation costs, and standardized mobile or in-vehicle availability of communication and GPS functionality.

Intelligent Vehicles and Traffic Control

We now discuss recent and future developments in connection with intelligent vehicles that can further improve the performance of traffic management and control systems by offering better and more accurate ways to collect traffic data and to apply traffic control measures.

Advanced Driver Assistance Systems

The increasing demand for safer passenger cars has stimulated the development of advanced driver assistance systems (ADAS). An ADAS is a control system that uses environment sensors to improve comfort and traffic safety by assisting the driver. Some examples of ADAS are cruise control, forward collision warning, lane departure warning, parking systems, and pre-crash systems for belt-pretensioning (Bishop 2005). Although traffic management is not the primary goal of ADAS, they can contribute to a better traffic performance (van Arem et al. 2006), either in a more passive way by avoiding incidents and by providing smoother traffic flows or in an active way by coordination and communication with neighboring vehicles and roadside infrastructure.

The increasing market penetration and use of ADAS and of other in-car navigation, telecommunication, and information systems offer an excellent

opportunity to implement a next level of traffic control and management, which shifts away from the road-side traffic management toward a vehicle-oriented traffic management. In this context both inter-vehicle management and road-side/vehicle traffic management and interaction can be considered. The goal is to use the additional measures and control handles offered by intelligent vehicles and to develop control and management methods to substantially improve traffic performance in terms of safety, throughput, reliability, environment, and robustness.

Some examples of new traffic control measures that are made possible by intelligent vehicles are cooperative adaptive cruise control (Vahidi and Eskandarian 2003) (allowing one to control intervehicle distances), intelligent speed adaptation (Carsten and Tate 2001) (allowing one to better and more dynamically control vehicle speeds), and route guidance (McDonald et al. 1995) (where the traffic control centers could on the one hand get data about planned routes and destinations, and on the other hand also send real-time information and control data to the on-board route planners, for instance, to warn about current and predicted congestion and possibly also to spread the traffic flows more evenly over the network).

These individual ADAS-based traffic control measures could be integrated with roadside traffic control measures such as ramp metering, traffic signals, lanes closures, and shoulder lane openings. The actual control strategy could then also make use of a model-based control approach such as MPC.

Cooperative Vehicle-Infrastructure Systems

The new intelligent-vehicle technologies allow communication and coordination between vehicles and the roadside infrastructure and among vehicles themselves. This results in cooperative vehicle-infrastructure systems, which can also be seen as a first step toward fully automated highway systems, which will be discussed below. CVIS (Cooperative Vehicle-Infrastructure Systems) (<http://www.cvisproject.org/>) is a European research project that aims to design, develop, and test technologies that allow communication

between the cars and with the roadside infrastructure, which improves road safety and efficiency, and reduces environmental impact. This project allows drivers to influence the traffic control system directly and also to get information about the quickest route to their destination, speed limits on the road, as well as warning messages via wireless technologies.

Automated Highway Systems

ADAS and cooperative vehicle-infrastructure systems can even be extended several steps further toward complete automation. Indeed, one approach to augment the throughput on highways is to implement a fully automated system called Automated Highway System (AHS) or Intelligent Vehicle/Highway System (IVHS) (Hedrick et al. 1994; Varaiya 1993), in which cars travel on the highway in platoons with small distances (say, 2 m) between vehicles within the platoon, and much larger distances (say, 30–60 m) between different platoons. Due to the very short intra-platoon distances this approach requires automated distance-keeping since human drivers cannot react fast enough to guarantee adequate safety. So in AHS every vehicle contains an automated system that can take over the driver's responsibilities in steering, braking, and throttle control. Due to the short spacing between the vehicles within the platoons, the throughput of the highway can increase, allowing it to carry as much as twice or three times as many vehicles as in the present situation. The other major advantages of the platooning system are increased safety and fuel efficiency. Safety is increased by the automation and close coordination between the vehicles and is enhanced by the small relative speed between the cars in the platoon. Because the cars in the platoon travel together at the same speed, a small distance apart, even high accelerations and decelerations cannot cause a severe collision between the cars (due to the small relative speeds). The short spacing between the vehicles also produces a significant reduction in aerodynamic drag for the vehicles, which leads to improvements in fuel economy and emissions reductions.

Automated platooning has been investigated very thoroughly within the PATH program

(<http://www.path.berkeley.edu/>; Shladover et al. 1991). Related programs are the Japanese Dolphin framework (Tsugawa et al. 2000) and the Auto21 Collaborative Driving System framework (<http://www.auto21.ca/index.php>; <http://www.damas.ift.ulaval.ca/projets/auto21/en/index.html>).

Although certain authors argue that only full automation can achieve significant capacity increases on highways and thus reduce the occurrences of traffic congestion (Varaiya 1993), AHS do not appear to be feasible on the short term. The AHS approach requires major investments to be made by both the traffic authority and the constructors and owners of the vehicles. Since few decisions are left to the driver, and since the AHS assumes almost complete control over the vehicles, which drive at high speeds and at short distances from each other, a strong psychological resistance to this traffic congestion policy is to be expected. Another important question is how the transition of the current highway system to an AHS-based system should occur, and – once it has been installed – what has to be done with vehicles that are not yet equipped for AHS. So before such systems can be implemented, many financial, legislative, political and organizational issues still have to be resolved (Fenton 1994).

Conclusion

In this entry we have presented an overview of freeway traffic control theory and practice. In this context we have discussed traffic measurements and estimation, individual traffic control measures, and the approaches behind them that relate the control signals to the given traffic situation. The trend of the ever-increasing traffic demands and the appearance of new control technologies have led to the new field of network-oriented traffic control systems. Although there have been many interesting publications about the theory and practice of integrated traffic control, several challenges remain, such as the integration of traffic state estimation and dynamic OD information in the control approaches.

The lively research in freeway traffic control shows that this field is still practically relevant and

theoretically challenging. Facing these challenges can be expected to lead to new freeway traffic control approaches in theory and practice resulting in higher freeway performance in terms of efficiency, reliability, safety and environmental effects. Furthermore, future developments in the field of in-car systems and advanced sensor technologies are expected to enable new traffic management approaches that may measure and control traffic in more detail and with higher performance.

Bibliography

Primary Literature

- Alessandri A, Di Febbraro A, Ferrara A, Punta E (1999) Nonlinear optimization for freeway control using variable-speed signaling. *IEEE Trans Veh Technol* 48(6):2042–2052
- Algers S, Bernauer E, Boero M, Breheret L, Di Taranto C, Dougherty M, Fox K, Gabard J-F (2000) SMARTTEST – final report for publication. Technical report, ITS, University of Leeds. <http://www.its.leeds.ac.uk/projects/smartest>. Accessed 22 July 2008
- André M, Hammarström U (2000) Driving speeds in Europe for pollutant emissions estimation. *Transp Res D* 5(5):321–335
- Åström KJ, Wittenmark B (1997) Computer controlled systems, 3rd edn. Prentice Hall, Upper Saddle River
- Banks JH (2005) Metering ramps to divert traffic around bottlenecks: some elementary theory. *Transp Res Rec* 1925:12–19
- Bell MGH (2000) A game theory approach to measuring the performance reliability of transport networks. *Transp Res B* 34(6):533–545
- Bellemans T (2003) Traffic control on motorways. PhD thesis, Katholieke Universiteit Leuven, Leuven. <ftp://ftp.esat.kuleuven.ac.be/pub/SISTA/bellemans/PhD/03-82.pdf>. Accessed 22 July 2008
- Bellemans T, De Schutter B, De Moor B (2006a) Model predictive control for ramp metering of motorway traffic: a case study. *Control Eng Pract* 14:757–767
- Bellemans T, De Schutter B, Wets G, De Moor B (2006b) Model predictive control for ramp metering combined with extended Kalman filter-based traffic state estimation. In: Proceedings of the 2006 I.E. intelligent transportation systems conference (ITSC 2006), Toronto, pp 406–411
- Bishop R (2005) Intelligent vehicle technology and trends. In: Walker J (ed) Artech House ITS library. Artech House, Boston
- Braess D (1968) Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12:258–268
- Branston D (1976) Models of single lane time headway distributions. *Transp Sci* 10:125–148
- Brownfield J, Graham A, Eveleigh H, Maunsell F, Ward H, Robertson S, Allsop R (2003) Congestion and accident risk. In: Technical report, Road Safety Research Report no. 44. Department for Transport, London
- Byon Y-J, Shalaby A, Abdulhai B (2006) Travel time collection and traffic monitoring via GPS technologies. In: Proceedings of the 2006 I.E. intelligent transportation systems conference (ITSC 2006), Toronto, pp 677–682
- Camacho EF, Bordons C (1995) Model predictive control in the process industry. Springer, Berlin
- Cambridge Systematics, Inc. (2001) Twin cities ramp meter evaluation – final report. Cambridge Systematics, Inc., Oakland. Prepared for the Minnesota Department of Transportation
- Carsten O, Tate F (2001) Intelligent speed adaptation: the best collision avoidance system? In: Proceedings of the 17th international technological conference on the enhanced safety of vehicles, Amsterdam, pp 1–10
- Carvell JD, Balke K Jr, Ullman J, Fitzpatrick K, Nowlin L, Brehmer C (1997) Freeway management handbook. Technical report. Federal Highway Administration, Department of Transport (FHWA, DOT), Washington, DC, Report No. FHWA-SA-97-064
- Castello P, Coelho C, Ninnio ED (1999) Traffic monitoring in motorways by real-time number plate recognition. In: Proceedings of the 10th international conference on image analysis and processing, Venice, pp 1129–1131
- Chaudhary NA, Messer CJ (2000) Ramp metering technology and practice. Technical report 2121-1. Texas Transportation Institute, The Texas A&M University System, College Station
- Chen OJ, Hotz AF, Ben-Akiva ME (1997) Development and evaluation of a dynamic ramp metering control model. In: Proceedings of the 8th IFAC/IFIP/IFORS symposium on transportation systems, Chania, June 1997, pp 1162–1168
- Chen A, Yang H, Lo HK, Tang WH (2002) Capacity reliability of a road network: an assessment methodology and numerical results. *Transp Res B* 36(3):225–252
- Chien C-C, Zhang Y, Ioannou PA (1997) Traffic density control for automated highway systems. *Automatica* 33(7):1273–1285
- Cools M, Moons E, Wets G (2007) Investigating the effect of holidays on daily traffic counts: a time series approach. *Transp Res Rec* 2019:22–31
- Cuena J, Hernández J, Molina M (1995) Knowledge-based models for adaptive traffic management systems. *Transp Res C* 3(5):311–337
- Daganzo CF (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B* 28B(4):269–287
- Daganzo CF (1996) The nature of freeway gridlock and how to prevent it. In: Lesort JB (ed) Proceedings of the 13th international symposium of transportation and traffic theory, Lyon, France. Emerald Group Publishing Limited, Bingley, UK
- Daganzo CF (1997) Fundamentals of transportation and traffic operations. Pergamon, Kidlington

- Di Febbraro A, Parisini T, Sacone S, Zoppoli R (2001) Neural approximations for feedback optimal control of freeway systems. *IEEE Trans Veh Technol* 50(1): 302–312
- Diakaki C, Papageorgiou M, McLean T (2000) Integrated traffic-responsive urban corridor control strategy in Glasgow, Scotland. *Transp Res Rec* 1727:101–111
- Elefteriadou L (1997) Freeway merging operations: a probabilistic approach. In: *Proceedings of the 8th international federation of automatic control (IFAC) symposium on transportation systems*, Chania, pp 1351–1356
- Federal Highway Administration, US Department of Transportation (1995a) Detection technology for IVHS – final report addendum. Technical report. Federal Highway Administration, US Department of Transportation, Washington, DC, Contract DTFH61-91-C-00076, Report No. FHWA-RD-96-109
- Federal Highway Administration, US Department of Transportation (1995b) Detection technology for IVHS – task L final report. Technical report. Federal Highway Administration, US Department of Transportation, Washington, DC, Contract DTFH61-91-C-00076, Report No. FHWA-RD-95-100
- Fenton RE (1994) IVHS/AHS: driving into the future. *IEEE Control Syst Mag* 14(6):13–20
- Gartner NH, Improti G (eds) (1995) *Urban traffic networks – dynamic flow modeling and control*. Springer, Berlin
- Gupta A, Maslanka VJ, Spring GS (1992) Development of prototype knowledge-based expert system for managing congestion on Massachusetts turnpike. *Transp Res Rec* 1358:60–66
- Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Transp Res Rec* 1320:91–98
- Hardman EJ (1996) Motorway speed control strategies using SISTM. In: *Proceedings of the 8th international conference on road traffic monitoring and control*. IEE conference publication, no. 422, London, 23–25 Apr 1996, pp 169–172
- Hasan M, Jha M, Ben-Akiva M (2002) Evaluation of ramp control algorithms using microscopic traffic simulation. *Transp Res C* 10:229–256
- Hedrick JK, Tomizuka M, Varaiya P (1994) Control issues in automated highway systems. *IEEE Control Syst Mag* 14(6):21–32
- Hegyi A, De Schutter B, Hellendoorn J (2005a) Model predictive control for optimal coordination of ramp metering and variable speed limits. *Transp Res C* 13(3):185–209
- Hegyi A, De Schutter B, Hellendoorn J (2005b) Optimal coordination of variable speed limits to suppress shock waves. *IEEE Trans Intell Transp Syst* 6(1):102–112
- Hegyi A, Girimonte D, Babuška R, De Schutter B (2006) A comparison of filter configurations for freeway traffic state estimation. In: *Proceedings of the international IEEE conference on intelligent transportation systems 2006*, Toronto, 17–20 Sept 2006, pp 1029–1034
- Helbing D (1997) *Verkehrsdynamik – Neue physikalische Modellierungskonzepte*. Springer, Berlin
- Hellinga B, Van Aerde M (1995) Examining the potential of using ramp metering as a component of an ATMS. *Transp Res Rec* 1494:75–83
- Hernández J, Cuenca J, Molina M (1999) Real-time traffic management through knowledge-based models: the TRYS approach. In: *ERUDIT tutorial on intelligent traffic management models*, Helsinki, 3 Aug 1999. <http://www.erudit.de/erudit/events/tc-c/tut990803.htm>. Accessed 23 July 2008
- Hoogendoorn SP (1997) Optimal control of dynamic route information panels. In: *Proceedings of the 4th world congress on intelligent transportation systems*, IFAC transportation systems, Chania, pp 399–404
- Hoogendoorn SP, Bovy PHL (2000) Continuum modelling of multiclass traffic flow. *Transp Res B* 34:123–146
- Hoogendoorn SP, Bovy PHL (2001) State-of-the-art of vehicular traffic flow modelling. *J Syst Control Eng Proc Inst Mech Eng I* 215(14):283–303
- Hotz A, Much C, Goblick T, Corbet E, Waxman A, Ashok K, Ben-Akiva M, Koutsopoulos H (1992) A distributed, hierarchical system architecture for advanced traffic management systems and advanced traffic information systems. In: *Proceedings of the second annual meeting of IVHS AMERICA*, Newport Beach
- <http://www.auto21.ca/index.php>. Accessed 23 July 2008
- <http://www.cvisproject.org/>. Accessed 23 July 2008
- <http://www.damas.ift.ulaval.ca/projets/auto21/en/index.html>. Accessed 23 July 2008
- <http://www.path.berkeley.edu/>. Accessed 23 July 2008
- Jacobson L, Henry K, Mehryar O (1989) Real-time metering algorithm for centralized control. *Transp Res Rec* 1232:17–26
- Jiang G, Gang L, Cai Z (2006) Impact of probe vehicles sample size on link travel time estimation. In: *Proceedings of the 2006 I.E. intelligent transportation systems conference (ITSC 2006)*, Toronto, pp 891–896
- Karimi A, Hegyi A, De Schutter B, Hellendoorn H, Middeham F (2004) Integration of dynamic route guidance and freeway ramp metering using model predictive control. In: *Proceedings of the 2004 American control conference (ACC 2004)*, Boston, 30 June–2 July 2004, pp 5533–5538
- Kell JH, Fullerton IJ, Mills MK (1990) *Traffic detector handbook*, Number FHWA-IP-90-002, 2nd edn. Federal Highway Administration, US Department of Transportation, Washington, DC
- Kerner BS (2002) Empirical features of congested patterns at highway bottlenecks. *Transp Res Rec* 1802:145–154
- Kerner BS (2004) *The physics of traffic*. Understanding complex systems. Springer, Berlin
- Kerner BS (2007a) Control of spatiotemporal congested traffic patterns at highway bottlenecks. *IEEE Trans Intell Transp Syst* 8(2):308–320
- Kerner BS (2007b) On-ramp metering based on three-phase traffic theory. *Traffic Eng Control* 48(1):28–35

- Kerner BS (2009) Introduction to modern traffic flow theory and control. Springer, Heidelberg/Dordrecht/London/New York
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory. *Physica A* 392:5261–5282
- Kerner BS (2015) Failure of classical traffic flow theories: a critical review. *Elektrotech Inf* 132(7):417–433
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Physica A* 450:700–747
- Kerner BS (2017) Breakdown minimization principle versus Wardrop's equilibria for dynamic traffic assignment and control in traffic and transportation networks: a critical mini-review. *Physica A* 466:626–662
- Kerner BS, Koller M, Klenov SL, Rehborn H, Leibel M (2015) The physics of empirical nuclei for spontaneous traffic breakdown in free flow at highway bottlenecks. *Physica A* 438:365–397
- Klein LA, Kelley MR, Mills MK (1997) Evaluation of overhead and in-ground vehicle detector technologies for traffic flow measurement. *J Test Eval* 25(2):215–224
- Klein LA, Mills MK, Gibson DRP (2006) Traffic detector handbook, Number FHWA-HRT-06-108, 3rd edn. Federal Highway Administration, US Department of Transportation, Washington, DC
- Kotsialos A, Papageorgiou M (2005) A hierarchical ramp metering control scheme for freeway networks. In: Proceedings of the American control conference, Portland, 8–10 June 2005, pp 2257–2262
- Kotsialos A, Papageorgiou M, Middelham F (2001) Optimal coordinated ramp metering with advanced motorway optimal control. In: Proceedings of the 80th annual meeting of the Transportation Research Board, vol 3125, Washington, DC
- Kotsialos A, Papageorgiou M, Mangeas M, Haj-Salem H (2002) Coordinated and integrated control of motorway networks via non-linear optimal control. *Transp Res C* 10(1):65–84
- Kraan M, van der Zijpp N, Tutert B, Vonk T, van Megen D (1999) Evaluating networkwide effects of variable message signs in the Netherlands. *Transp Res Rec* 1689:60–67
- Kühne RD (1991) Freeway control using a dynamic traffic flow model and vehicle reidentification techniques. *Transp Res Rec* 1320:251–259
- Lee HY, Lee H-W, Kim D (1999) Empirical phase diagram of traffic flow on highways with on-ramps. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) Traffic and granular flow '99. Springer, Berlin
- Lenz H, Sollacher R, Lang M (2001) Standing waves and the influence of speed limits. In: Proceedings of the European control conference 2001, Porto, pp 1228–1232
- Levinson D, Zhang L (2006) Ramp meters on trial: evidence from the twin cities metering holiday. *Transp Res A* 40A:810–828
- Lewis FL (1992) Applied optimal control and estimation. Prentice Hall, Upper Saddle River
- Li PY, Horowitz R, Alvarez L, Frankel J, Robertson AM (1995) Traffic flow stabilization. In: Proceedings of the American control conference, Seattle, June 1995, pp 144–149
- Lighthill MJ, Whitham GB (1955) On kinematic waves, II. A theory of traffic flow on long crowded roads. *Proc R Soc* 229A(1178):317–345
- Lin P-W, Chang G-L (2007) A generalized model and solution algorithm for estimation of the dynamic freeway origin-destination matrix. *Transp Res B* 41B:554–572
- Little R, Rubin D (1987) Handbook of transportation science. Wiley, New York
- Lo HK, Luo XW, Siu BWY (2006) Degradable transport network: travel time budget of travelers with heterogeneous risk aversion. *Transp Res B* 40(9):792–806
- Maciejowski JM (2002) Predictive control with constraints. Prentice Hall, Harlow
- Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419(1):1–64
- May AD (1990) Traffic flow fundamentals. Prentice Hall, Englewood Cliffs
- McDonald M, Hounsell NB, Njoze SR (1995) Strategies for route guidance systems taking account of driver response. In: Proceedings of 6th vehicle navigation and information systems conference, Seattle, July 1995, pp 328–333
- Middelham F (1999) A synthetic study of the network effects of ramp metering. Technical report. Transport Research Centre (AVV), Dutch Ministry of Transport, Public Works and Water Management, Rotterdam
- Middelham F (2003) State of practice in dynamic traffic management in The Netherlands. In: Proceedings of the 10th IFAC symposium on control in transportation systems (CTS 2003), Tokyo, Aug 2003
- Middleton D, Gopalakrishna D, Raman M (2002) Advances in traffic data collection and management. Technical report BAT-02-006. Texas Transportation Institute, Cambridge Systematics Inc., Washington, DC. http://www.itsdocs.fhwa.dot.gov/IPODOCS/REPTS_TE/13766.html. Accessed 23 July 2008
- Molina M, Hernández J, Cuenca J (1998) A structure of problem-solving methods for real-time decision support in traffic control. *Int J Hum Comput Stud* 49:577–600
- Papageorgiou M (1980) A new approach to time-of-day control based on a dynamic freeway traffic model. *Transp Res B* 14B:349–360
- Papageorgiou M (1983) A hierarchical control system for freeway traffic. *Transp Res B* 17B(3):251–261
- Papageorgiou M (ed) (1991) Concise encyclopedia of traffic & transportation systems. Pergamon, Oxford, UK
- Papageorgiou M (1998) Some remarks on macroscopic traffic flow modelling. *Transp Res A* 32(5):323–329
- Papageorgiou M, Kotsialos A (2000) Freeway ramp metering: an overview. In: Proceedings of the 3rd annual IEEE conference on intelligent transportation systems (ITSC 2000), Dearborn, Oct 2000, pp 228–239

- Papageorgiou M, Kotsialos A (2002) Freeway ramp metering: an overview. *IEEE Trans Intell Transp Syst* 3(4):271–280
- Papageorgiou M, Messmer A (1991) Dynamic network traffic assignment and route guidance via feedback regulation. *Transp Res Rec* 1306:49–58
- Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990a) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: part I: modelling. *Transp Res A* 24A(5):345–359
- Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990b) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: part II: coordinated on-ramp metering. *Transp Res A* 24A(5):361–370
- Papageorgiou M, Hadj-Salem H, Blosseville J-M (1991) ALINEA: a local feedback control law for on-ramp metering. *Transp Res Rec* 1320:58–64
- Papageorgiou M, Hadj-Salem H, Middelham F (1997) ALINEA local ramp metering – summary of field results. *Transp Res Rec* 1603:90–98
- Papageorgiou M, Wang Y, Kosmatopoulos E, Papamichail I (2007) ALINEA maximizes motorway throughput – an answer to flawed criticism. *Traffic Eng Control* 48(6):271–276
- Payne HJ (1971) Models of freeway traffic and control. *Simul Counc Proc* 1:51–61
- Rees I (1995) Orbital decongestant. *Highways* 63(5):17–18
- Rees T, Harbord B, Dixon C, Abou-Rhame N (2004) Speed-control and incident detection on the M25 controlled motorway (summary of results 1995–2002). Technical report PPR033, TRL (UK's Transport Research Laboratory), Wokingham
- Richards PI (1956) Shock waves on the highway. *Oper Res* 4:42–51
- Ritchie SG (1990) A knowledge-based decision support architecture for advanced traffic management. *Transp Res A* 24A(1):27–37
- Schik P (2003) Einfluss von Streckenbeeinflussungsanlagen auf die Kapazität von Autobahnabschnitten sowie die Stabilität des Verkehrsflusses. PhD thesis, Universität Stuttgart
- Shladover SE, Desoer CA, Hedrick JK, Tomizuka M, Walrand J, Zhang WB, McMahon DH, Peng H, Sheikholesham S, McKeown N (1991) Automatic vehicle control developments in the PATH program. *IEEE Trans Veh Technol* 40(1):114–130
- Sisiopiku VP (2001) Variable speed control: technologies and practice. In: *Proceedings of the 11th annual meeting of ITS America, Miami*, pp 1–11
- Slotine JE, Li W (1991) *Applied nonlinear control*. Prentice Hall, Englewood Cliffs
- Smith BL, Zhang H, Fontaine MD, Green MW (2004) Wireless location technology based traffic monitoring: critical assessment and evaluation of an early-generation system. *J Transp Eng* 130(5):576–584
- Smulders S (1990) Control of freeway traffic flow by variable speed signs. *Transp Res B* 24B(2):111–132
- Smulders SA, Helleman DE (1998) Variable speed control: state-of-the-art and synthesis. In: *Road transport information and control, number 454 in conference publication, IEE*, 21–23 Apr 1998, pp 155–159
- Soriguera F, Thorson L, Robuste F (2007) Travel time measurement using toll infrastructure. In: *Proceedings of the 86th annual meeting of the Transportation Research Board, CDROM paper 07-1389.pdf*, Washington, DC
- Taylor MAP (1999a) Dense network traffic models, travel time reliability and traffic management. I: general introduction. *J Adv Transp* 33(2):218–233
- Taylor MAP (1999b) Dense network traffic models, travel time reliability and traffic management. II: application to network reliability. *J Adv Transp* 33(2):235–251
- Taylor CJ, Young PC, Chotai A, Whittaker J (1998) Non-minimal state space approach to multivariable ramp metering control of motorway bottlenecks. *IEE Proc Control Theory Appl* 146(6):568–574
- Taylor MAP, Woolley JE, Zito R (2000) Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transp Res C* 8C:257–285
- Transportation Research Board (1998) Managing speed: review of current practice for setting and enforcing speed limits. In: *TRB special report 254, Transportation Research Board, Committee for Guidance on Setting and Enforcing Speed Limits*. National Academy Press, Washington, DC
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulation. *Phys Rev E* 62:1805–1824
- Tsugawa S, Kato S, Tokuda K, Matsui T, Fujii H (2000) An architecture for cooperative driving of automated vehicles. In: *Proceedings of the IEEE intelligent transportation symposium, Dearborn*, pp 422–427
- Vahidi A, Eskandarian A (2003) Research advances in intelligent collision avoidance and adaptive cruise control. *IEEE Trans Intell Transp Syst* 4(3):143–153
- van Arem B, van Driel CJG, Visser R (2006) The impacts of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Trans Intell Transp Syst* 4:429–436
- van den Berg M, Bellemans T, De Schutter B, De Moor B, Hellendoorn J (2005) Control of traffic with anticipative ramp metering. In: *Proceedings of the 84th annual meeting of the Transportation Research Board, Washington, DC, 2005 CDROM paper 05-0252*, pp 166–174
- van den Hoogen E, Smulders S (1994) Control by variable speed signs: results of the Dutch experiment. In: *Proceedings of the 7th international conference on road traffic monitoring and control, IEE conference publication no. 391, London, 26–28 Apr 1994*, pp 145–149
- Varaiya P (1993) Smart cars on smart roads: problems of control. *IEEE Trans Autom Control* 38(2):195–207
- Wang Y, Papageorgiou M (2005) Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transp Res B* 39(2):141–167

- Wang Y, Papageorgiou M, Messmer A (2003) A predictive feedback routing control strategy for freeway network traffic. In: Proceedings of the 82nd annual meeting of the Transportation Research Board, Washington, DC, Jan 2003
- Wang Y, Papageorgiou M, Messmer A (2006) RENAISSANCE – a unified macroscopic model-based approach to real-time freeway network traffic surveillance. *Transp Res C* 14C:190–212
- Wattleworth JA (1965) Peak-period analysis and control of a freeway system. *Highw Res Rec* 157:1–21
- Wilkie JK (1997) Using variable speed limit signs to mitigate speed differentials upstream of reduced flow locations. Technical report. Department of Civil Engineering, Texas A&M University, College Station. Prepared for CVEN 677 Advanced Surface Transportation Systems, Report No. SWUTC/97/72840-00003-2
- Williams B (1996) Highway control. *IEE Rev* 42(5):191–194
- Wolf DE (1999) Cellular automata for traffic simulations. *Physica A* 263:438–451
- Zhang L (2007) Traffic diversion effect of ramp metering at the individual and system levels. In: Proceedings of the 86th annual meeting of the Transportation Research Board, Washington, DC, Jan 2007, CDROM paper 07-2087
- Zhang H, Ritchie SG (1994) Real-time decision-support system for freeway management and control. *J Comput Civ Eng* 8(1):35–51

Books and Reviews

We refer the interested reader to the following references in the various fields that have been discussed in this entry:

Control: General introduction (Åström and Wittenmark 1997), optimal control (Lewis 1992), model predictive control (Camacho and Bordons 1995; Maciejowski 2002), nonlinear control (Slotine and Li 1991)

Traffic flow modeling: General overviews (Hoogendoorn and Bovy 2000, 2001), cell transmission model (Daganzo 1994), Kerner's three-phase theory (Kerner 2004), microscopic simulation models (Algers et al. 2000), cellular automata (Maerivoet and De Moor 2005)

Ramp metering: Overviews of ramp metering strategies (Carvell et al. 1997; Papageorgiou and Kotsialos 2000), field test and simulation studies (Hasan et al. 2002)

Speed limit systems: Overviews of practical speed limit systems (Sisiopiku 2001; Wilkie 1997)

Intelligent vehicles: Overview (Bishop 2005)

Sensor technologies: Overviews (Federal Highway Administration, US Department of Transportation 1995a, b; Klein et al. 1997, 2006)

Modeling Approaches to Traffic Breakdown

Boris S. Kerner
Physics of Transport and Traffic, University
Duisburg-Essen, Duisburg, Germany

Article Outline

Glossary
Definition of the Subject and its Importance
Introduction
Free and Congested Traffic in Empirical Data
Three Traffic Phases
Probabilistic Empirical Features of Traffic
Breakdown ($F \rightarrow S$ Transition)
Empirical Proof of Nucleation Nature of Traffic
Breakdown ($F \rightarrow S$ Transition)
The Basic Assumption of Three-Phase Traffic
Theory
Empirical Nucleation Nature of Traffic
Breakdown as Origin of the Infinity of
Highway Capacities at Any Time Instant
Failure of Classical Understanding of Stochastic
Highway Capacity
Failure of Classical Modeling Approaches to
Traffic Breakdown at Highway Bottlenecks
Main Prediction of Three-Phase Theory
Competition of Driver Speed Adaptation with
Driver Over-Acceleration in Three-Phase
Theory
Basic Requirement for a Three-Phase Traffic
Flow Model
Kerner-Klenov Stochastic Microscopic Three-
Phase Traffic Flow Model
Evaluation of on-Ramp Metering with Three-
Phase Theory
Conclusions
Future Directions
Bibliography

Glossary

Bottleneck Traffic breakdown occurs mostly at road bottlenecks. Just as defects and impurities are important for phase transitions in complex spatially distributed systems of various nature, so are bottlenecks in vehicular traffic. A road bottleneck can be a result of roadworks, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, traffic signal, etc.

Congested Traffic Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is possible in free flow.

$F \rightarrow S$ transition In all known observations, traffic breakdown at a highway bottleneck is a phase transition from the free flow phase to synchronized flow phase ($F \rightarrow S$ transition). The empirical traffic breakdown ($F \rightarrow S$ transition) exhibits the nucleation nature. The empirical nucleation nature of traffic breakdown is explained by the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck. The terms *traffic breakdown* and an $F \rightarrow S$ transition are synonyms related to the same phenomenon of the onset of congestion in free flow.

Free Flow Free flow is usually observed, when the vehicle density in traffic is small enough. The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function.

Fundamental Diagram of Traffic Flow The fundamental diagram is a relationship between the flow rate and density in vehicle traffic. Because the flow rate is the product of the average speed and density, the fundamental diagram is associated with a relationship between these traffic variables. In accordance with an obvious result of traffic measurements, on average the speed

decreases when the density increases. Thus, in the flow–density plane, the fundamental diagram should pass through the origin (when the density is zero so is the flow rate) and should have at least one maximum. The fundamental diagram gives also a connection between the average space gap (net distance) between vehicles and the average speed.

Highway Capacity of Free Flow at Bottleneck

Highway capacity of free flow at a bottleneck is limited by traffic breakdown at the bottleneck. Any flow rate in metastable free flow at the bottleneck at which traffic breakdown ($F \rightarrow S$ transition) can be induced at a highway bottleneck is equal to one of the highway capacities of free flow at the bottleneck. There are the infinite number of such flow rates: *At any time instant*, there are the infinite number of highway capacities of free flow at the bottleneck that are between a minimum highway capacity and a maximum highway capacity. The existence of an infinite number of highway capacities at any time instant means that highway capacity is stochastic.

Induced Traffic Breakdown (Induced $F \rightarrow S$ transition) at Bottleneck

An induced traffic breakdown (induced $F \rightarrow S$ transition) in metastable free flow at a bottleneck is traffic breakdown that is induced at the bottleneck by the propagation of a moving spatiotemporal congested traffic pattern. This congested pattern has occurred *earlier* than the time instant of traffic breakdown at the bottleneck and at a *different* road location (e.g., at another bottleneck) than the bottleneck location. When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck. The empirical induced $F \rightarrow S$ transition at the bottleneck proves the empirical nucleation nature of traffic breakdown at the bottleneck. The empirical nucleation nature of traffic breakdown at a highway bottleneck can be considered the empirical fundamental of transportation science.

Moving Traffic Jam A moving traffic jam (moving jam for short) is a localized structure of large vehicle density spatially limited by two jam fronts. Within the downstream jam front, vehicles accelerate escaping from the jam;

within the upstream jam front, vehicles slow down approaching the jam. The jam as a whole structure propagates upstream in traffic flow. Within the jam (between the jam fronts), vehicle density is large and speed is very low (sometimes as low as zero). A sequence of moving jams is also called “stop-and-go” traffic or oscillations in congested traffic.

Spontaneous Traffic Breakdown (Spontaneous $F \rightarrow S$ transition) at Bottleneck

If before traffic breakdown occurs at a highway bottleneck there is free flow at the bottleneck as well as upstream and downstream in a neighborhood of the bottleneck, then traffic breakdown is called spontaneous traffic breakdown (spontaneous $F \rightarrow S$ transition) at the bottleneck. In empirical data, a spontaneous traffic breakdown occurs, if a nucleus required for traffic breakdown appears spontaneously at the bottleneck.

Steady States of Traffic Flow Steady states of traffic flow are hypothetical states of homogeneous (in time and space) traffic flow of identical vehicles (and identical drivers) in which all vehicles move with the same time-independent speed and have the same space gaps between vehicles.

Synchronized Flow Traffic Phase In three-phase traffic theory, the following definition of the synchronized flow phase [S] in congested traffic based on measured traffic data is made. The downstream front of synchronized flow does not exhibit the characteristic feature of wide moving jams; specifically, the latter front is often fixed at a road bottleneck.

Three-Phase Traffic Theory In the three-phase traffic theory (three-phase theory for short), besides the free flow phase there are two phases in congested traffic, the synchronized flow (S) and wide moving jam (J) phases. The synchronized flow and wide moving jam phases in congested traffic are defined through the use of empirical spatiotemporal criteria (definitions) [S] and [J], respectively. The three-phase theory is a qualitative theory that explains empirical traffic breakdown and resulting empirical congested traffic patterns based on the criteria [S] and [J] for traffic

phases as well as on hypotheses of this theory. The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown at highway bottlenecks. In the three-phase theory, the empirical nucleation nature of traffic breakdown at a highway bottleneck is explained by the metastability of free flow with respect to an $F \rightarrow S$ transition at the bottleneck.

Traffic Breakdown In empirical observations, when density in free flow increases and becomes large enough, the phenomenon of the onset of congestion traffic is observed in free flow: The average speed decreases sharply to a lower speed in congested traffic. This speed breakdown observed during the onset of congestion is called the breakdown phenomenon or traffic breakdown.

Traffic Flow Model A traffic flow model is devoted to the explanation and simulation of traffic flow phenomena, which are observed in measured data of real traffic flow, and to the prediction of new traffic flow phenomena that could be found in real traffic flow. First of all, a traffic flow model should explain and predict empirical (measured) features of traffic breakdown.

Wide Moving Jam Traffic Phase In three-phase traffic theory, the following definition of the wide moving jam phase [J] in congested traffic based on measured traffic data is made. A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. This is the characteristic feature of the wide moving jam phase.

Definition of the Subject and its Importance

Traffic breakdown is a transition from free flow to congested traffic in a traffic or transportation network. Traffic breakdown occurs usually at network bottlenecks.

Users of traffic and transportation networks expect that through the application of traffic control, dynamic traffic assignment, cooperative driving systems, as well as other intelligent

transportation systems (ITS), traffic breakdown in a network can be prevented. This is because congested traffic resulting from traffic breakdown causes a considerably increase in travel time, fuel consumption, CO₂ emission, as well as other travel costs. Therefore, any traffic and transportation theory applied for the development of automated driving vehicles, reliable methods of dynamic traffic assignment, and control should be consistent with empirical features of traffic breakdown at a network bottleneck.

The most important empirical feature of traffic breakdown in a network is the nucleation nature of traffic breakdown found in real field traffic data. The term *nucleation nature* of traffic breakdown means that traffic breakdown occurs in a metastable free flow. The metastability of free flow is as follows. There can be many speed (density, flow rate) disturbances in free flow. Amplitudes of the disturbances can be very different. When a disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown occurs. Such a disturbance resulting in the breakdown is called *nucleus* for the breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays; as a result, no traffic breakdown occurs.

- The nucleation nature of traffic breakdown at a network bottleneck is an empirical fundamental of transportation science.

Unfortunately, generally accepted classical traffic and transportation theories, which have had a great impact on the understanding of many empirical traffic phenomena, have nevertheless failed by their applications in the real world. Even several decades of a very intensive effort to improve and validate network optimization and control models based on the classical traffic and transportation theories have had no success. Indeed, there can be found no examples where on-line implementations of the network optimization models based on these classical traffic and transportation theories could reduce congestion in real traffic and transportation networks.

This failure of classical traffic and transportation theories can be explained as follows: The classical traffic and transportation theories are not consistent with the nucleation nature of traffic

breakdown. The nucleation nature of empirical traffic breakdown was understood only during last 20 years. In contrast, the generally accepted fundamentals and methodologies of traffic and transportation theory were introduced in the 1950s–1960s. Thus, the scientists whose ideas led to the classical traffic and transportation theories could not know the empirical nucleation nature of traffic breakdown. Respectively, in 1996–2002 the author introduced the three-phase traffic theory.

- The three-phase theory is the theoretical fundamental of transportation science that explains the empirical nucleation nature of traffic breakdown.

The classical traffic and transportation theories that failed by their applications in the real world are currently the methodologies of teaching programs in most universities and the subject of publications in most transportation research journals and scientific conferences. Consequently, some of article's objectives are as follows:

1. We prove the empirical nucleation nature of traffic breakdown in networks.
2. We discuss the origin of the failure of classical modeling approaches to traffic breakdown.
3. We show that the development of traffic flow modeling approaches in the framework of the three-phase theory is needed for the reliable evaluation of ITS-applications like traffic control, dynamic traffic assignment, the effect of automated driving vehicles on mixed traffic flow, etc.

Modeling approaches for the explanation of traffic breakdown include traffic flow models based on a diverse variety of methods of complexity and systems science. This is explained by an extremely complex nonlinear character of real vehicular traffic. Moreover, there are links between spatiotemporal phenomena associated with traffic breakdown in vehicular traffic and traffic phenomena observed in many other complex systems like pedestrian traffic, which is responsible for the complexity of crowd and evacuation dynamics, and air traffic. For this reason,

the understanding of traffic breakdown in vehicular traffic is also very important for the further progress in understanding of many other many-particle complex systems.

Introduction

In this entry, we limit a consideration of approaches to the modeling of traffic breakdown at a highway bottleneck. Readers can find features of traffic breakdown at traffic signal in city traffic in chapter “► [Autonomous Driving in the Framework of Three-Phase Traffic Theory](#)” of the book (Kerner 2017a) as well as in the article (Kerner 2017b) of this Encyclopedia. A highway bottleneck can be caused by on- and off-ramps, reduction of highway lines, road works, etc. Traffic breakdown at the bottleneck is a local phase transition from free flow (F) to congested traffic whose downstream front is fixed at the bottleneck (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990). In the three-phase theory (Kerner 2004a), congested traffic that downstream front is fixed at the bottleneck is called synchronized flow. Therefore, empirical traffic breakdown is a transition from free flow to synchronized flow at the bottleneck ($F \rightarrow S$ transition).

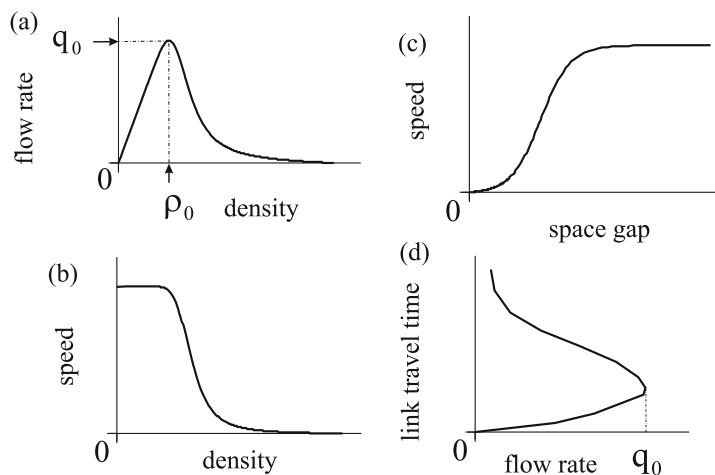
Empirical studies show that during traffic breakdown vehicle speed sharply decreases. In contrast, after traffic breakdown has occurred, the flow rate can remain as large as in an initial free flow. For this reason, traffic breakdown is also called *speed drop* or *speed breakdown* (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990).

Earlier traffic flow models, which claim to explain and predict traffic breakdown at road bottlenecks, are reviewed in (Bellomo et al. 2002; Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Whitham 1974; Wiedemann 1974; Wolf 1999). One of the main hypotheses of

these classical traffic flow models is the hypothesis about a theoretical fundamental diagram (Fig. 1a) (Bellomo et al. 2002; Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Whitham 1974; Wiedemann 1974; Wolf 1999). The fundamental diagram for a traffic flow model means that in all these traffic flow models, it is assumed that for a given time-independent speed of the preceding vehicle, there is a single desired (or optimal) space gap for the following vehicle in a hypothetical acceleration less (deceleration less) and noiseless model limit case. The theoretical fundamental diagram of traffic flow is associated with the obvious results of the observations of traffic: The larger the vehicle density, the lower is the average speed in road traffic (Bellomo et al. 2002; Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999).

The classical traffic flow models can be classified into two main classes. The first model class refers to the classic LWR (Lighthill-Whitham-Richards) model introduced in 1955–1956 (Lighthill and Whitham 1955; Richards 1956) (see also references in Chowdhury et al. 2000; Daganzo 1997; Gartner et al. 1997; Haight 1963; Helbing 2001; Leutzbach 1988; May 1990; Treiber and Kesting 2013; Whitham 1974). Examples for this model class are the cell-transmission models of Daganzo (Daganzo 1994, 1995). The basic idea of the LWR-model is that maximum flow rate (denoted q_0 in Fig. 1a), associated with the maximum point at the fundamental diagram, determines capacity of free flow at a bottleneck, i.e., the bottleneck capacity. Thus, if the flow rate upstream of the bottleneck exceeds the bottleneck capacity, then traffic breakdown should occur.

The second model class refers to the classic GM (General Motors) model of Herman, Gazis, Rothery, Montroll, Chandler, and Potts introduced in 1959–1961 (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959). The basic idea of the GM-model approach is as follows. Beginning at a critical density, there is instability of steady model states at the fundamental diagram.



Modeling Approaches to Traffic Breakdown, Fig. 1 Qualitative example of fundamental diagram: (a) Flow-density relationship (fundamental diagram). (b–d) The speed-density (b), speed-space-gap (c), and link-travel-time-flow (d) relationships associated with (a) (see, e.g., Bellomo et al. 2002; Chowdhury et al. 2000;

Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999)

This model instability that should explain traffic breakdown is associated with a finite value of driver reaction time. The instability can qualitatively be explained, if we assume that in a homogeneous traffic flow a vehicle decelerates unexpectedly, which causes a local speed decrease (speed disturbance) in the flow. Due to the driver reaction time, the following vehicle can decelerate stronger than it would be necessary to avoid collision. This effect is called the over-deceleration effect (or human over-reaction). As a result of over-deceleration, the speed of this vehicle can become lower than the speed of the preceding vehicle. The same over-deceleration effect can occur for many other following vehicles that can lead to the growth of the initial speed disturbance, i.e., to traffic flow instability. Examples for the GM model class are the optimal velocity (OV) models by Newell, by Whitham, and by Bando et al., the Payne macroscopic model and its variants, the Wiedemann psychophysical traffic flow model and its variants, the Gipps-model, the NaSch (Nagel-Schreckenberg) cellular automata (CA) model and its variants, the Krauß-model, the IDM-model (Intelligent Driver Model) of Treiber et al. as well as a huge number of other traffic flow models (see below and references in Berg and Woods 2001; Brockfeld et al. 2003, 2005; Ceder 1999; Chowdhury et al. 2000; Cremer 1979; Fukui et al. 2003; Gartner et al. 1997; Haight 1963; Helbing 2001; Lesort 1996; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999).

The most important task of traffic flow modeling is to explain and simulate empirical features of traffic breakdown at network bottlenecks. This is because congested traffic resulting from traffic breakdown in a traffic network causes a considerably increase in travel time, fuel consumption, CO₂ emission, as well as other travel costs. Therefore, any traffic flow model applied for the analysis of the performance of automated driving vehicles, reliable methods of dynamic traffic assignment, and control as well as any other ITS-applications (ITS – intelligent transportation systems) should be consistent with the empirical

features of traffic breakdown at a network bottleneck. As we have already explained in the review article (Kerner 2017b), the empirical traffic breakdown at the bottleneck is a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) that exhibits the nucleation nature. The empirical nucleation nature of traffic breakdown is an empirical fundamental of transportation science.

The nucleation nature of traffic breakdown means that traffic breakdown occurs in *metastable* free flow at the bottleneck with respect to an $F \rightarrow S$ transition: Small enough speed disturbances in the metastable free flow at the bottleneck decay while causing no breakdown. However, when a large enough local disturbance occurs in the metastable free flow at the bottleneck, then traffic breakdown does occur. Such a large local speed disturbance in free flow at the bottleneck can be considered a nucleus for traffic breakdown.

The empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck has been understood only about 20 years ago (Kerner 1998a, b, 1999a). It turns out that *none* of earlier traffic flow models (Bellomo et al. 2002; Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999) can explain and show the empirical nucleation nature of the $F \rightarrow S$ transition at the bottleneck.

This explains why traffic flow simulations in the frameworks of the LWR model and/or traffic flow models of the GM model class *cannot* be used for reliable and valid analysis of ITS performance in the real world.

In the recent book (Kerner 2017a), we have explained this failure of classical traffic theories. We have also shown that for a reliable analysis of traffic flow phenomena, a traffic flow model in the framework of the three-phase traffic theory (Kerner 1998a, b, c, 1999a, b, c, 2000, 2001, 2002a, b, 2004a, 2009c) is needed.

However, the book (Kerner 2017a) is about 650 pages long. Additionally to a detailed critical consideration of traffic breakdown at highway

bottlenecks, in the book many other very different fields of traffic and transportation science like a theory of traffic breakdown at traffic signals in city traffic, the effect of automated driving vehicles on traffic flow, traffic control, and dynamic traffic assignment in traffic networks have been studied.

In contrast to the book (Kerner 2017a), this review article is devoted to a brief critical discussion of modeling approaches to traffic breakdown at highway bottlenecks only. Moreover, in the book (Kerner 2017a), there are more than 1200 references to works of several generations of traffic and transportation researchers. For this reason, in this review article, we limit referring only to some original works as well as to some reviews and books needed for the understanding of results and conclusions of this brief critical review.

The article is organized as follows. In sections “Free and Congested Traffic in Empirical Data,” “Three Traffic Phases,” “Probabilistic Empirical Features of Traffic Breakdown ($F \rightarrow S$ Transition),” and “Empirical Proof of Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition),” empirical features of highway traffic related to the subject of this article are discussed. Firstly, we consider the well-known empirical definition of congested traffic (section “Free and Congested Traffic in Empirical Data”) as well as empirical definitions of two different traffic phases in congested traffic – the synchronized flow phase and the wide moving jam phase made in the three-phase traffic theory (section “Three Traffic Phases”). Then, some well-known empirical probabilistic features of traffic breakdown and the related classical understanding of stochastic highway capacity will be discussed (section “Probabilistic Empirical Features of Traffic Breakdown ($F \rightarrow S$ Transition)”). In section “Empirical Proof of Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition),” we consider the empirical proof of the nucleation nature of traffic breakdown at highway bottlenecks. We will find that traffic breakdown is a transition from free flow to synchronized flow ($F \rightarrow S$ transition) occurring in metastable free flow at a bottleneck with respect to the $F \rightarrow S$ transition. We emphasize that in accordance with explanations made in

the book (Kerner 2017a), the nucleation nature of traffic breakdown at highway bottlenecks can be considered the empirical fundamental of transportation science. The basic assumption of the three-phase theory is considered in section “The Basic Assumption of Three-Phase Traffic Theory.” In section “Empirical Nucleation Nature of Traffic Breakdown as Origin of the Infinity of Highway Capacities at Any Time Instant,” we explain that and why the empirical nucleation nature of traffic breakdown leads to the conclusion of the three-phase theory that at any time instant there is the infinity of highway capacities. Then, we show that the classical understanding of stochastic highway capacity contradicts the empirical nucleation nature of traffic breakdown (section “Failure of Classical Understanding of Stochastic Highway Capacity”). In section “Failure of Classical Modeling Approaches to Traffic Breakdown at Highway Bottlenecks,” based on a consideration of earlier classical traffic flow models, we explain that and why these models cannot show the empirical nucleation nature of the $F \rightarrow S$ transition (traffic breakdown) at highway bottlenecks. The main prediction of the three-phase theory is discussed in section “Main Prediction of Three-Phase Theory.” A competition of driver speed adaptation with driver over-acceleration that explains the metastability of free flow with respect to an $F \rightarrow S$ transition at highway bottlenecks is the subject of section “Competition of Driver Speed Adaptation with Driver Over-Acceleration in Three-Phase Theory.” The basic requirement for a three-phase traffic flow model is considered in section “Basic Requirement for a Three-Phase Traffic Flow Model.” In section “Kerner-Klenov Stochastic Microscopic Three-Phase Traffic Flow Model,” we discuss the Kerner-Klenov microscopic stochastic three-phase traffic flow model that shows and explains the empirical nucleation nature of the $F \rightarrow S$ transition at highway bottlenecks. To illustrate the application of this three-phase traffic flow model for the analysis of the reliability of ITS-applications, in section “ANCONA On-Ramp Metering” we compare the performance of ANCONA on-ramp metering method (Kerner 2005, 2007a, b, c, d) associated with the understanding of highway capacity in the

framework of the three-phase theory with a well-known ALINEA on-ramp metering method of Papageorgiou et al. (1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) that is based on the classical understanding of highway capacity.

Free and Congested Traffic in Empirical Data

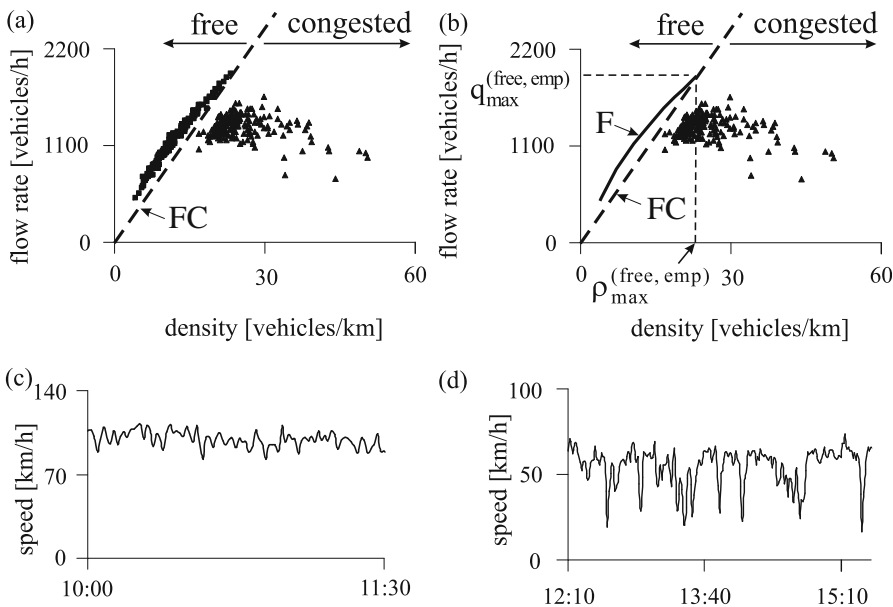
At small enough vehicle density, interactions between vehicles in free flow are negligible. Therefore, vehicles have an opportunity move with their desired maximum speeds (if this speed is not restricted by road conditions or traffic regulations). When the density increases, vehicle interaction cannot be neglected any more. As a result of vehicle interaction in free flow, the average speed decreases with increase in density (measured data left of the dashed line FC in Fig. 2a).

This common vehicle behavior in free flow leads to the *fundamental diagram for free flow*,

i.e., a certain curve with a positive slope that gives a relationship between the flow rate and density. This relationship is associated with the averaging of measured data shown left of the dashed line FC in Fig. 2a to one average flow rate for each density (curve F in Fig. 2b) (e.g., Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Helbing 2001; Helbing et al. 2000; Hoogendoorn et al. 2005; Mahmassani 2005; May 1990; Newell 1982; Prigogine and Herman 1971; Whitham 1974; Wiedemann 1974).

The empirical fundamental diagram for free flow is cut off at some empirical limit (maximum) point of free flow (denoted $(\rho_{\max}^{(\text{free, emp})}, q_{\max}^{(\text{free, emp})})$ in Fig. 2b) (e.g., Hall et al. 1992; May 1990). At this limit point of free flow, the average vehicle speed has a minimum possible value for free flow denoted by $v_{\max}^{(\text{free, emp})}$. Thus, empirical points of free flow as well as the associated fundamental diagram lie to the left of the dashed line FC in the flow–density plane whose slope equals this minimum possible speed in free flow.

When the vehicle density is large enough, traffic is usually in a congested state (e.g., Chowdhury



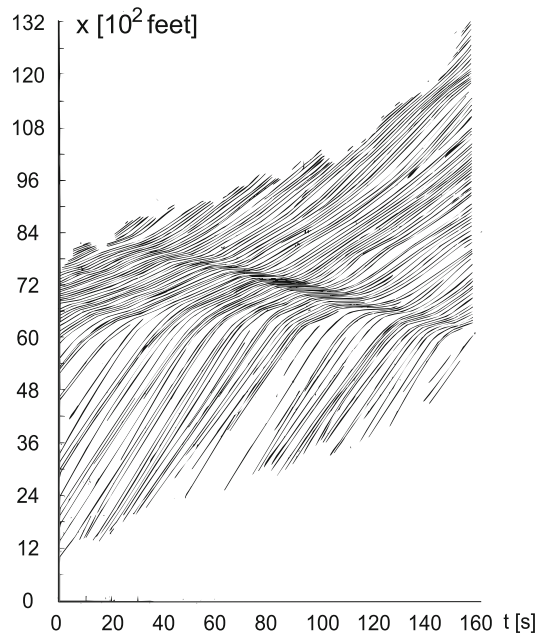
Modeling Approaches to Traffic Breakdown, Fig. 2 Free flow and congested traffic in empirical traffic data (e.g., Koshi et al. 1983). (a) Data for free flow (points left of the dashed line FC) and for congested traffic (points right of the dashed line FC) measured by a road detector.

(b) The fundamental diagram for free flow (curve F) and the same measured data for congested traffic as those in (a). (c, d) Vehicle speed in free flow (c) and congested traffic (d), related to points left and right of the line FC in (a), respectively. 1-min average data

et al. 2000; Cremer 1979; Daganzo 1997; Fukui et al. 2003; Helbing 2001; Helbing et al. 2000; Hoogendoorn et al. 2005; Mahmassani 2005; May 1990; Newell 1982; Prigogine and Herman 1971; Schadschneider et al. 2007; Schreckenberg and Wolf 1998; Taylor 2002; Whitham 1974; Wiedemann 1974). We define traffic phenomena in terms of their empirical features found in measured data rather than their genesis. Often, congested traffic is defined as a traffic state, which merely results from a freeway bottleneck as the inevitable consequence of an upstream flow that exceeds the bottleneck capacity. In this definition, it seems obvious to assume that congested flow is not self-organized, but rigidly controlled by external constraints. However, it has been shown that a variety of self-organized processes are responsible for features of real traffic congestion (Kerner 2004a). For these reasons, we use the following *definition* of congested traffic.

Congested traffic states is defined as complementary to states of free flow (e.g., Koshi et al. 1983). It has already been noted that empirical points related to free flow can be approximately presented by a curve with a positive slope in the flow–density plane (Fig. 2b). Congested traffic will be defined, therefore, as a state of traffic where the average vehicle speed is *lower* than the minimum possible speed in free flow $v_{\max}^{(\text{free}, \text{emp})}$, which is related to the limit point $(\rho_{\max}^{(\text{free}, \text{emp})}, q_{\max}^{(\text{free}, \text{emp})})$ in free flow (Fig. 2b). Thus, empirical points of congested traffic lie to the right of the dashed line FC in the flow–density plane whose slope equals the minimum possible speed in free flow. In this definition of congested traffic, nothing is said about the origin of the congestion. It is related to the empirical fact that congested traffic can occur spontaneously at a bottleneck.

In congested traffic, the “stop-and-go” phenomenon is observed, i.e., a sequence of different moving traffic jams can appear (Edie and Foote 1958, 1960; Edie 1961; Edie et al. 1980; Koshi 2003; Koshi et al. 1983; Treiterer and Taylor 1966; Treiterer and Myers 1974; Treiterer 1975). Moving jams have been studied empirically by many authors, in particular, in classic empirical works by Edie et al. (1980; Edie and Foote 1958, 1960; Edie 1961), Treiterer et al. (Treiterer and



Modeling Approaches to Traffic Breakdown, Fig. 3 An empirical moving jam: traffic dynamics derived from aerial photography (Adapted from Treiterer 1975)

Taylor 1966; Treiterer and Myers 1974; Treiterer 1975) (Fig. 3) and Koshi et al. (Koshi 2003; Koshi et al. 1983) (see other references to empirical and theoretical results of analyses of moving jam, e.g., in (Chowdhury et al. 2000; Helbing 2001; Kerner 2004a; Nagatani 2002; Nagel et al. 2003; Rehborn and Klenov 2009; Rehborn et al. 2017; Treiber and Kesting 2013).

Three Traffic Phases

A traffic phase is a traffic state considered in space and time that possesses specific *empirical spatio-temporal* features. These features are specific (unique) only to this traffic phase. Note that a traffic state is characterized by a certain set of statistical properties of traffic variables. Examples of traffic variables are the flow rate q [vehicles/h], vehicle speed v [km/h], vehicle space gap that is also called as net distance or space headway g [m], time headway [s] that is also called time space or net time gap between vehicles, and vehicle density ρ [vehicles/km].

Based on investigations of empirical features of traffic breakdown at highway bottlenecks as well as resulting empirical congested spatiotemporal patterns measured on different highways over many days and years (see references in Kerner 2004a), the three-phase traffic theory introduced by the author in 1996–1999 suggests that besides the free flow traffic phase there are two other traffic phases in congested traffic: synchronized flow and wide moving jam.

Traffic flow consists of free flow and congested traffic. Congested traffic consists of the synchronized flow phase and the wide moving jam phase. Thus, there are three traffic phases in three-phase traffic theory:

Free flow

Synchronized flow

Wide moving jam

The synchronized flow and wide moving jam phases in congested traffic are defined through the use of the following *empirical (objective) criteria for the traffic phases in congested traffic*. These empirical criteria are related to some characteristic spatiotemporal empirical features of the traffic phases (Kerner 2004a):

[J] The wide moving jam traffic phase is defined as follows. A wide moving jam is a moving jam that maintains the mean velocity of the downstream *front* of the jam, v_g , as the jam propagates. Vehicles accelerate within the downstream jam front from low speed states (sometimes as low as zero) inside the jam to higher speeds downstream of the jam. On average a wide moving jam maintains the mean velocity of the downstream jam front, even as it propagates through other (possibly very complex) traffic states or freeway bottlenecks. This is a characteristic feature of the wide moving jam traffic phase.

[S] The synchronized flow traffic phase is defined as follows. In contrast to the wide moving jam traffic phase, the downstream front of the synchronized flow traffic phase does *not* maintain the mean velocity of the downstream front. In

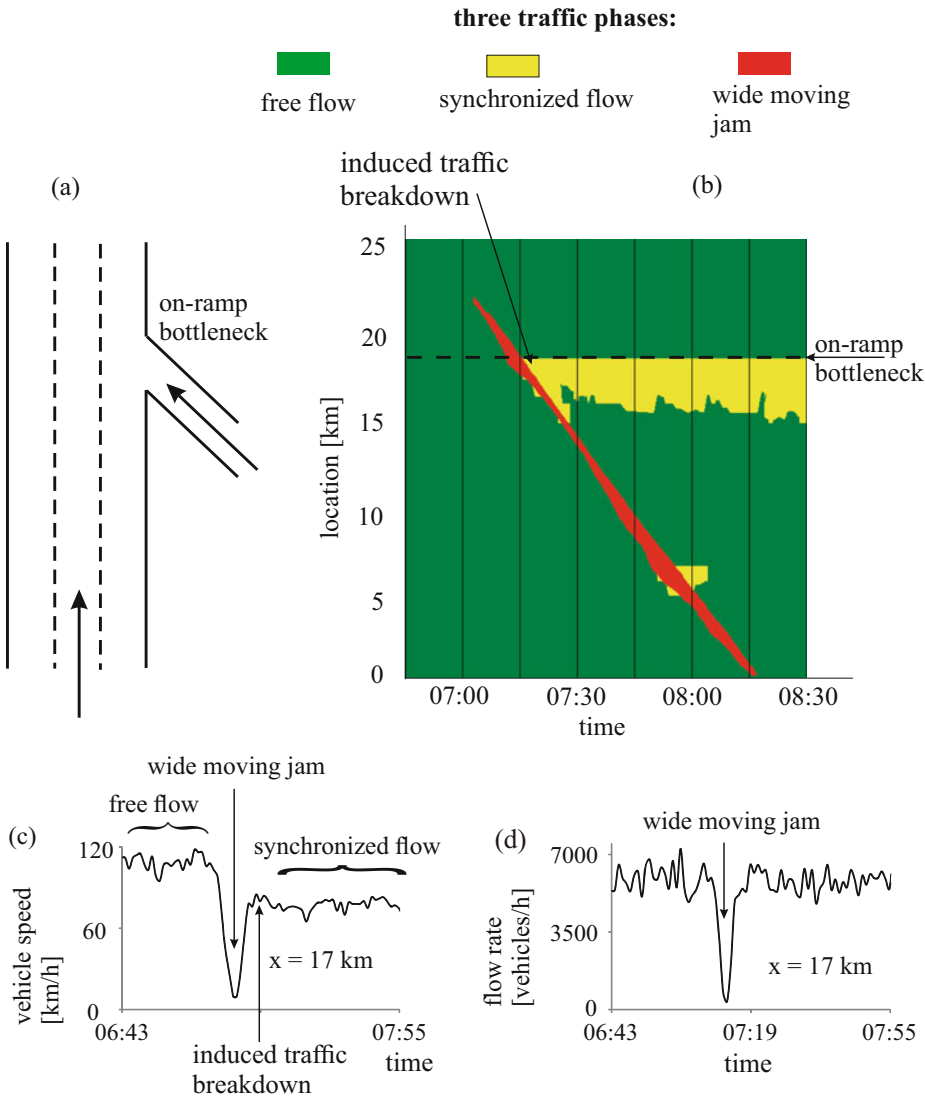
particular, the downstream front of synchronized flow is often *fixed* at a freeway bottleneck. In other words, the synchronized flow traffic phase does not show aforementioned characteristic feature of the wide moving jam traffic phase.

The downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow, vehicles accelerate from lower speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front.

Thus, the definitions of the traffic phases in congested traffic are made via the spatiotemporal empirical criteria [J] and [S]. The definitions [J] and [S] are associated with dynamic behavior of the downstream front of these phases, while a wide moving jam or synchronized flow propagates through other traffic states or bottlenecks.

The definitions [J] and [S] of the traffic phases in congested traffic mean that if a congested traffic state is not related to the wide moving jam phase, then with certainty the state is associated with the synchronized flow phase. This is because congested traffic can be either within the synchronized flow phase or within the wide moving jam phase. In other words, if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase. In particular, it must be stressed that a moving jam can be associated either with a wide moving jam or with synchronized flow. In accordance with the phase definitions [J] and [S], this depends on the dynamic behavior of the downstream jam front, while the jam propagates through other traffic states or bottlenecks.

An empirical example of application of the objective criteria (definitions) for the traffic phases in congested traffic [J] and [S] is shown in Fig. 4. In Fig. 4, a moving jam phase propagating through the location of an on-ramp bottleneck *induces* the synchronized flow phase at this bottleneck (this induced phase transition is labeled “induced traffic breakdown” in Fig. 4b). Indeed, in accordance with the definition [S], the downstream front of synchronized



Modeling Approaches to Traffic Breakdown, Fig. 4 Empirical definitions of three traffic phases based on example of empirical induced traffic breakdown at on-ramp bottleneck that is induced by wide moving jam propagation: (a) Sketch of section of three-lane highway in Germany with an on-ramp bottleneck. (b) Average data of vehicle speed in space and time measured with road

detectors installed along road section in (a) (1-min averaging interval); data are presented in space and time with averaging method described in section C.2 of (Kerner et al. 2013b). (c, d) Speed (c) and total flow rate (d) over time at the location of the bottleneck $x = 17$ km. Real field traffic data measured on three-lane freeway A5-South in Germany on June 23, 1998 (Adapted from Kerner 2004a)

flow is fixed at the bottleneck. In contrast, the moving jam propagates through this bottleneck with the mean velocity of the downstream jam front remaining unchanged. Thus, in accordance with the definition [J], this moving jam is associated with the

wide moving jam phase. Synchronized flow is self-sustaining for a very long time (more than an hour) upstream of the bottleneck. The possibility of the induced $F \rightarrow S$ transition means that an $F \rightarrow S$ transition exhibits nucleation character that is a

characteristic feature of metastable free flow with respect to the $F \rightarrow S$ transition.

Probabilistic Empirical Features of Traffic Breakdown ($F \rightarrow S$ Transition)

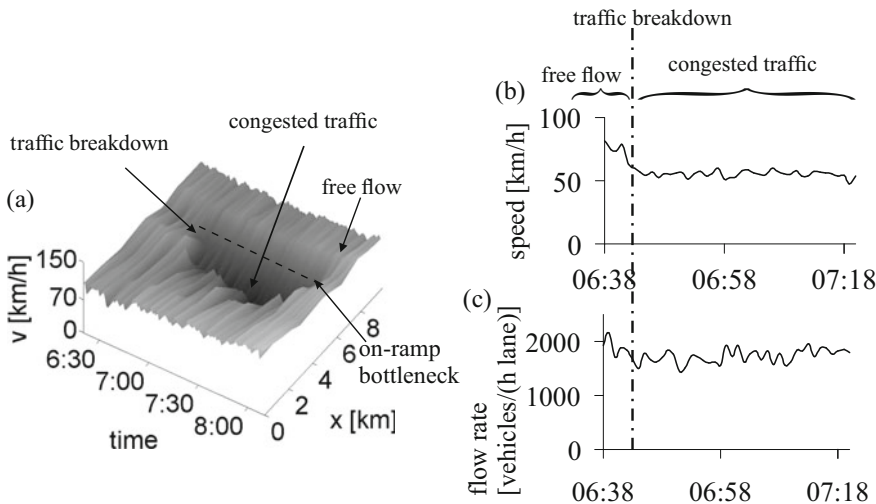
Traffic breakdown is almost daily observed in traffic networks of any industrial country of the world. As already emphasized in the article introduction, traffic breakdown is a transition from free flow to congested traffic. Traffic breakdown with resulting traffic congestion occurs usually at a road bottleneck. Therefore, highway capacity of free flow is limited by traffic breakdown at the bottleneck (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990 and references there as well as references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)).

Beginning from the classical work by Greenshields et al. (1935), a great effort has been made by many scientific groups to understand empirical features of traffic breakdown (see, e.g.,

Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990 as well as references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)).

Empirical studies show that during traffic breakdown vehicle speed sharply decreases (Fig. 5a, b). In contrast, after traffic breakdown has occurred, the flow rate can remain as large as in an initial free flow (traffic breakdown is labeled by “traffic breakdown” in Fig. 5c). For this reason, traffic breakdown is also called *speed drop* or *speed breakdown* (Banks 1989; Banks 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990).

Traffic breakdown at a highway bottleneck is a local phase transition from free flow (F) to congested traffic whose downstream front is usually fixed at the bottleneck location (Fig. 5a) (Banks 1989, 1990; Bovy 1998; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; Highway



Modeling Approaches to Traffic Breakdown, Fig. 5 Empirical example of traffic breakdown at on-ramp bottleneck: (a) Averaged vehicle speed in space and time (1-min averaging interval). (b, c) Vehicle speed (b) and averaged flow rate (per road lane) (c) measured on freeway in a neighborhood of the on-ramp bottleneck at detector location $x = 6.4$ km. Real measured traffic data of road detectors adapted from (Kerner 2004a) that was measured

on three-lane freeway A5-South in Germany on March 26, 1996; the data exhibits common features that are qualitatively the same as those measured on a variety of highways in different countries (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990 as well as references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)).

Capacity Manual 2000, 2010; May 1990). As mentioned in Section “Three Traffic Phases,” in the three-phase theory such congested traffic is called synchronized flow (S) (Kerner 2004a). In other words, using the terminology of the three-phase theory, traffic breakdown is a transition from free flow to synchronized flow ($F \rightarrow S$ transition) (Kerner 2004a). However, it should be emphasized that as long as nucleation features of synchronized flow are not discussed (this discussion will be done in section “Empirical Proof of Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition),” the term *synchronized flow* is nothing more as only the definition of congested traffic whose downstream front is fixed at the bottleneck.

There is also a well-known hysteresis phenomenon associated with traffic breakdown and a return transition to free flow (Fig. 6) (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; Highway Capacity Manual 2000, 2010; May 1990 as well as references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)).

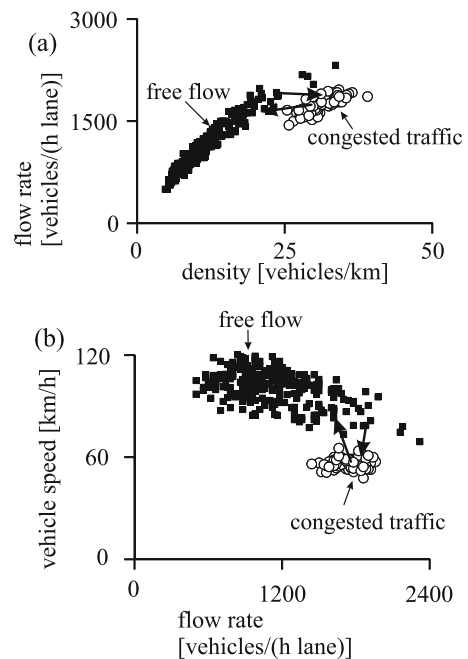
Stochastic (Probabilistic) Behavior of Traffic Breakdown

In 1995, Elefteriadou et al. found that traffic breakdown at a highway bottleneck has a stochastic (probabilistic) behavior (Elefteriadou 2014; Elefteriadou et al. 1995). This means the following: At a given flow rate in free flow at the bottleneck traffic breakdown can occur, but it should not necessarily occur. Thus, on one day traffic breakdown occurs, however, on another day at the same flow rates traffic breakdown is not observed.

In 1998, Persaud et al. (1998) found that probability $P^{(B)}$ of traffic breakdown at the bottleneck is a sharp increasing function of the flow rate in free flow in the vicinity of the point, where the breakdown probability $P^{(B)} = 1$ (Fig. 7). This behavior of the breakdown probability $P^{(B)}$ is qualitatively independent of the averaging time interval T_{av} for the flow rate. However, the increase in this averaging time interval leads to a decrease in the flow rate at which $P^{(B)} = 1$ (curves $T_{av} = 1$ min and 10 min in Fig. 7). One of the reasons for this behavior of the flow-rate function

of the breakdown probability $P^{(B)}$ is that in empirical data the flow rate in free flow usually changes considerably over time. As a result, in 10 min averaged data (Fig. 7), there can be points of 1 min averaged data where the flow rate is considerably higher than the average flow rate in 10 min averaged data (Persaud et al. 1998).

Growing flow-rate functions of the probability for traffic breakdown at on-ramp bottlenecks has also been found for freeways in the USA by Lorenz and Elefteriadou (2000) as well as for 5 min averaged traffic data measured on German

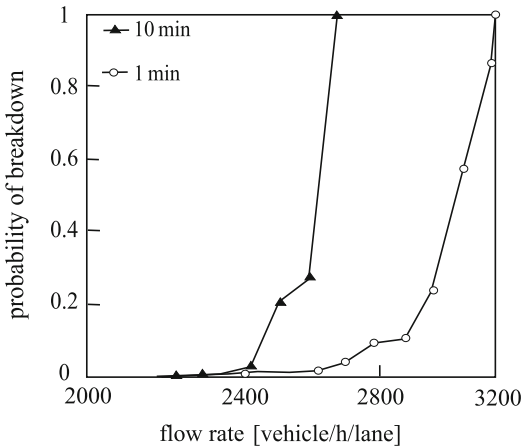


Modeling Approaches to Traffic Breakdown, Fig. 6 Typical empirical example of hysteresis phenomenon (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; Highway Capacity Manual 2000, 2010; May 1990 as well as references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)) presented in the flow–density (a) and speed–flow planes (b). This empirical example of hysteresis phenomenon, which is caused by traffic breakdown ($F \rightarrow S$ transition) and return transition to free flow ($S \rightarrow F$ transition), is related to traffic data measured in a neighborhood of the on-ramp bottleneck at detector location $x = 6.4$ km within traffic pattern shown in Fig. 5a. Black points – free flow, white circles – congested traffic (synchronized flow)

freeways by Brilon et al. (2005a; b; Brilon and Zurlinden 2004). The results are qualitatively similar to the findings of Persaud et al. (1998).

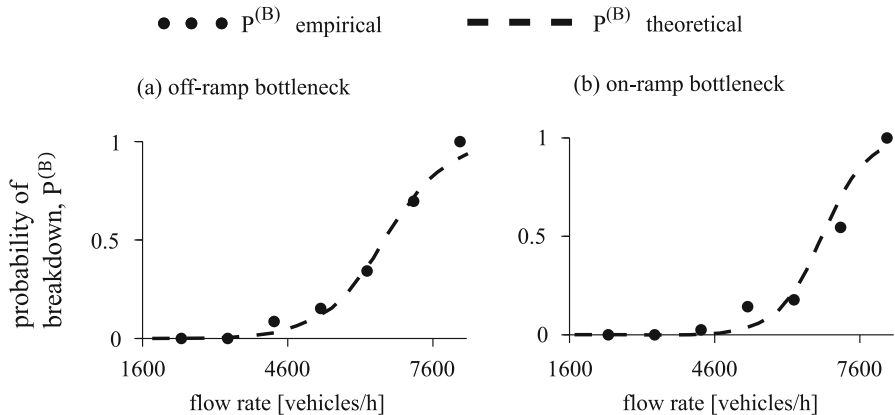
A growing character of the empirical flow-rate dependence of the probability of traffic

breakdown at a highway bottleneck discovered by Persaud et al. (1998) and further studied by Lorenz and Elefteriadou (2000) as well as by Brilon et al. (2005a, b; Brilon and Zurlinden 2004) has also been confirmed in our later empirical studies of traffic breakdown at off- and on-ramp bottlenecks with the use of 1-min averaged data measured on German highways (black points in Fig. 8) (Kerner et al. 2015). The empirical probabilities of traffic breakdown are well fitted with the theoretical flow-rate dependence of the breakdown probability (dashed curves in Fig. 8) that was revealed in 2002 by Kerner et al. (2002) in the framework of the three-phase theory (see formula (7) of section “The Basic Assumption of Three-Phase Traffic Theory”).



Modeling Approaches to Traffic Breakdown, Fig. 7 Probability for traffic breakdown at on-ramp bottleneck for two different averaging time intervals for traffic variables $T_{av} = 1$ and 10 min (Adapted from Persaud et al. 1998)

- The empirical probability of traffic breakdown at highway bottlenecks is a growing flow-rate function (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Lorenz and Elefteriadou 2000; Persaud et al. 1998) (see also references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)).



Modeling Approaches to Traffic Breakdown, Fig. 8 Empirical flow-rate dependence of the probability of traffic breakdown $P^{(B)}$ measured at an off-ramp bottleneck (a) and an on-ramp bottleneck (b): (a) Empirical breakdown probability (black points) was found from a study of traffic data in which traffic breakdown at the off-ramp bottleneck was observed on 89 different days. (b) Empirical breakdown probability (black points) at the on-ramp bottleneck was found from a study of traffic

data in which traffic breakdown was observed on 56 different days. Empirical breakdown probabilities (black points) are related to real field traffic data measured by road detectors installed along a section of three-lane freeway A5-South. Dashed curves are the theoretical flow-rate dependencies of the probability of traffic breakdown $P^{(B)}$ related to formula (7) of section “The Basic Assumption of Three-Phase Traffic Theory” (Adapted from Kerner et al. 2015)

Classical Understanding of Stochastic Highway Capacity

As above mentioned, highway capacity of free flow is limited by traffic breakdown at the bottleneck (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990 and references there as well as references in chapter “► Breakdown in Traffic Networks” of the book (Kerner 2017a)).

In the classical definition of stochastic highway capacity, it is assumed that at *each given time instant* there is only one particular value of stochastic highway capacity. As long as no traffic breakdown occurs at the bottleneck, we do not know highway capacity because it is a stochastic value. It is assumed that the particular value of stochastic highway capacity is equal to the flow rate in free flow at which traffic breakdown occurs at the bottleneck (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; Highway Capacity Manual 2000, 2010; May 1990). To explain the stochastic (probabilistic) behavior of traffic breakdown at a highway bottleneck (Elefteriadou et al. 1995), Brilon has introduced the following concept for stochastic highway capacity (Brilon et al. 2005a, b; Brilon and Zurlinden 2004).

In accordance with the above-mentioned classical understanding of stochastic highway capacity, Brilon’s stochastic highway capacity is equal to the flow rate in an initially free flow at the bottleneck at which traffic breakdown is observed at the bottleneck. At any time instant, there is a particular value of stochastic capacity of free flow at the bottleneck. However, as long as free flow is observed at the bottleneck, this particular value of stochastic capacity cannot be measured. Therefore, stochastic capacity is defined through a capacity distribution function $F_C^{(B)}(q_{\text{sum}})$, where q_{sum} is the flow rate at a highway bottleneck (Brilon et al. 2005a, b; Brilon and Zurlinden 2004):

$$F_C^{(B)}(q_{\text{sum}}) = p(C \leq q_{\text{sum}}), \quad (1)$$

where $p(C \leq q_{\text{sum}})$ is the probability that stochastic highway capacity C is equal to or smaller than the flow rate q_{sum} .

Thus, the basic theoretical assumption of the classical understanding of stochastic highway capacity is that traffic breakdown is observed at a time instant t at which the flow rate q_{sum} reaches the capacity C . This means that the flow rate function of the probability of traffic breakdown should be determined by the capacity distribution function (Brilon et al. 2005a, b; Brilon and Zurlinden 2004):

$$P^{(B)}(q_{\text{sum}}) = F_C^{(B)}(q_{\text{sum}}). \quad (2)$$

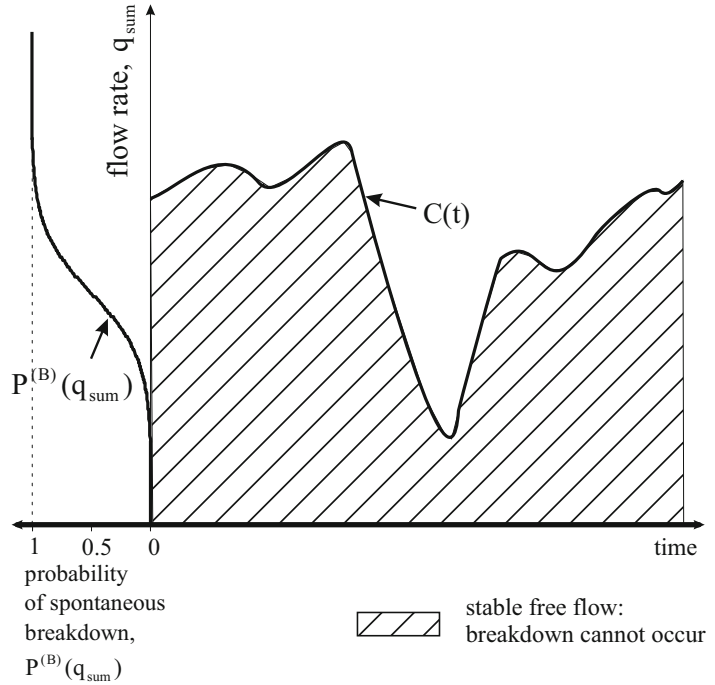
It must be noted that in accordance with empirical results presented in Fig. 7 (Persaud et al. 1998), the breakdown probability $P^{(B)}(q_{\text{sum}})$ is an increasing function of the flow rate (left in Fig. 9). Thus, this flow rate function $P^{(B)}(q_{\text{sum}})$ is the empirical fact (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Lorenz and Elefteriadou 2000; Persaud et al. 1998). However, Eq. 2 is a theoretical hypothesis only. This is because in contrast with the breakdown probability function $P^{(B)}(q_{\text{sum}})$, the capacity distribution function $F_C^{(B)}(q_{\text{sum}})$ cannot be measured. This understanding of stochastic capacity of free flow at a bottleneck, which is currently well accepted in the community of traffic and transportation engineers (Elefteriadou 2014; Elefteriadou et al. 2014), is illustrated in Fig. 9.

In Fig. 9, we show a qualitative hypothetical fragment of the time-dependence of stochastic capacity C over time t (Fig. 9, right). Left in Fig. 9, we show a qualitative flow rate dependence of the probability of spontaneous traffic breakdown $P^{(B)}(q_{\text{sum}})$. It is often assumed that a stochastic behavior of highway capacity is associated with a stochastic change in traffic parameters over time (Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990). Examples of the traffic parameters, which can indeed be stochastic time-functions in real traffic, are weather, mean driver’s characteristics (e.g., mean driver reaction time), share of long vehicles, etc.

In accordance with the definition of stochastic capacity (1), (2), no traffic breakdown can occur,

Modeling Approaches to Traffic Breakdown,

Fig. 9 Qualitative explanation of Brilon's stochastic highway capacity of free flow at a highway bottleneck (right figure). Left figure – a qualitative flow-rate function of the probability of the spontaneous breakdown $P^{(B)}(q_{\text{sum}})$



when the time dependence of the flow rate is given by a hypothetical time dependence $q_{\text{sum}}(t) = q_{\text{sum}, 1}(t)$ (Fig. 10a). This is because at all-time instants $q_{\text{sum}, 1}(t) < C(t)$.

In contrast, for another hypothetical time dependence $q_{\text{sum}}(t) = q_{\text{sum}, 2}(t)$ traffic breakdown should occur at time instant $t = t_1$ at which this flow rate is equal to the capacity value: $q_{\text{sum}, 2}(t_1) = C(t_1)$ (Fig. 10b).

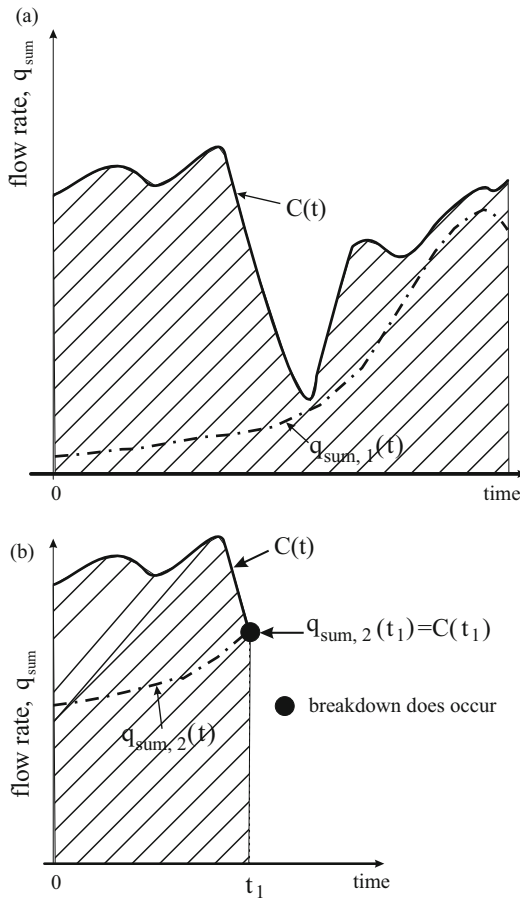
In other words, the classical understanding of stochastic capacity can be explained as follows: At a given time instant, no traffic breakdown can occur at a highway bottleneck if the flow rate in free flow at the bottleneck at the time instant is smaller than the value of the capacity at this time instant.

The basic importance of the words “at a given time instant” in the capacity definition is as follows: Because Brilon's stochastic capacity $C(t)$ changes stochastically over time (Fig. 9, right), clearly that at a given time instant traffic breakdown can occur at the flow rate that is smaller than the value of the stochastic capacity was at another time instant (Fig. 10b).

Summary of Achievements of Empirical Studies of Stochastic Highway Capacity

The achievements of empirical studies of traffic breakdown at road bottlenecks (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990) can be summarized as follows:

1. Traffic breakdown at a highway bottleneck is a local phase transition from free flow (F) to congested traffic whose downstream front is usually fixed at the bottleneck location. In the three-phase theory, such congested traffic has been called “synchronized flow” (S) (Kerner 2004a).
2. During empirical traffic breakdown, the vehicle speed decreases sharply. The flow rate in the emergent congested traffic, which is observed after the breakdown at the bottleneck location, can be as large as the flow rate has been in an initially free flow at the bottleneck before the breakdown has occurred (see, e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al.



Modeling Approaches to Traffic Breakdown, Fig. 10 Qualitative explanation of traffic breakdown with the use of Brilon's stochastic highway capacity of free flow at a highway bottleneck. The fragment of the hypothetical time-function of Brilon's stochastic highway capacity $C(t)$ is taken from Fig. 9 (Adapted from Kerner 2016)

1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990).

- Empirical breakdown phenomenon exhibits the probabilistic nature: At the same highway bottleneck traffic breakdown occurs at very different flow rates on different days of data observations as discovered firstly by Elefteriadou et al. (1995) and confirmed later in many empirical observations (see, e.g., Brilon et al. 2005a; Brilon and Zurlinden 2004; Lorenz and Elefteriadou 2000; Persaud et al. 1998 and references in the book by Elefteriadou 2014).

- The probability of empirical traffic breakdown is an increasing flow rate function (Fig. 7), as firstly discovered in 1998 by Persaud et al. (1998). Later this important empirical result has been confirmed in studies of real field traffic data measured in different countries (Brilon et al. 2005a; Brilon and Zurlinden 2004; Kerner et al. 2015; Lorenz and Elefteriadou 2000).
- There is a well-known hysteresis phenomenon associated with traffic breakdown and a return transition to free flow (e.g., Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990) (Fig. 6).

However, in section “**Empirical Proof of Nucleation Nature of Traffic Breakdown (F→S Transition)**” we will explain that the sole knowledge of the above-mentioned features of empirical traffic breakdown at highway bottlenecks revealed in (Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990) is *not* sufficient to disclose the physical nature of traffic breakdown and associated stochastic highway capacity.

Indeed, in section “**Empirical Proof of Nucleation Nature of Traffic Breakdown (F→S Transition)**” we will show that the classical understanding of stochastic highway capacity (section “**Classical Understanding of Stochastic Highway Capacity**”) that at any given time instant there is a particular value of highway capacity of free flow at a bottleneck (Banks 1989, 1990; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990) is invalid for real traffic.

Empirical Proof of Nucleation Nature of Traffic Breakdown (F → S Transition)

In observations of traffic breakdown (F → S transition) at a bottleneck in real field traffic data, we distinguish two different cases: (i) empirical *spontaneous* traffic breakdown (spontaneous F → S transition) and (ii) empirical *induced* traffic breakdown (induced F → S transition) (Kerner 2004a, 2009c):

- If before traffic breakdown occurs at the bottleneck, *free flow* exists at the bottleneck as well as upstream and downstream in a neighborhood of the bottleneck, then such an empirical traffic breakdown at the bottleneck is called empirical spontaneous traffic breakdown.
- An empirical induced traffic breakdown at the bottleneck is traffic breakdown induced by the propagation of a spatiotemporal *congested* traffic pattern. This congested pattern has occurred *earlier* than the time instant of traffic breakdown at the bottleneck and at a *different* road location (for example at another bottleneck) than the bottleneck location. When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck.

An example of empirical spontaneous traffic breakdown has been presented in Fig. 5a. Indeed, in this case before traffic breakdown occurs at a bottleneck, there is free flow at the bottleneck as well as upstream and downstream in a neighborhood of the on-ramp bottleneck. Other two examples of empirical spontaneous traffic breakdown, which will be considered below in more details, are shown in Fig. 11.

Two examples of empirical induced traffic breakdown are shown in Figs. 4 and 12. In the first example, a wide moving jam propagating through an on-ramp bottleneck *induces* the synchronized flow phase ($F \rightarrow S$ transition) at this bottleneck (“induced traffic breakdown” in Fig. 4).

In the second example, a moving synchronized flow pattern (MSP) that has initially emerged at the downstream off-ramp bottleneck propagates upstream. When the MSP reaches an upstream on-ramp bottleneck (labeled by “on-ramp bottleneck 1” in Fig. 12), the MSP induced the synchronized flow phase ($F \rightarrow S$ transition) at this on-ramp bottleneck (induced traffic breakdown is labeled by “ind” in Fig. 12b and by “induced traffic breakdown” in Fig. 12c, g).

In both examples of empirical induced traffic breakdown (induced $F \rightarrow S$ transition) (Figs. 4 and 12), the downstream front of resulting congested traffic is fixed at the bottleneck, i.e., in accordance with the definition [S] this congested traffic is

associated with the synchronized flow phase. Synchronized flow is self-sustaining for a very long time (more than an hour) upstream of the bottleneck.

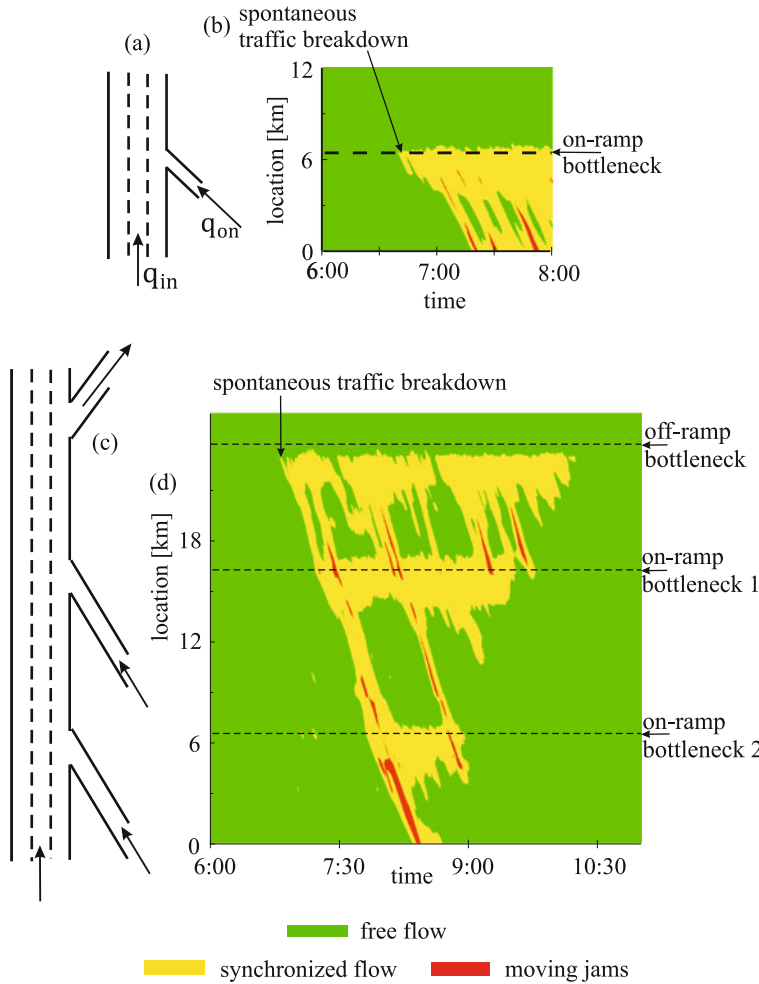
There are some differences between these two examples of empirical induced traffic breakdown (Figs. 4 and 12). In the case, when empirical induced traffic breakdown has been caused by the propagation of the wide moving jam through the bottleneck, the flow rate within the wide moving jam shown in Fig. 4d is very small. Contrarily, when empirical induced traffic breakdown has been caused by the propagation of the MSP to the bottleneck, the flow rate within the MSP shown in Fig. 12h is almost as large as in the surrounded free flow. Another difference between the wide moving jam shown in Fig. 4 and the MSP shown in Fig. 12 is as follows: The wide moving jam propagates through the bottleneck while maintaining the mean velocity of its downstream front (Fig. 4b). In contrast, the MSP is caught at the bottleneck (Fig. 12b). This effect has been called “catch effect.”

In both examples of empirical induced traffic breakdown (Figs. 4 and 12) after induced traffic breakdown has occurred at the bottleneck, characteristics of synchronized flow that has emerged at the bottleneck due to the breakdown are qualitatively the same as those found in the case of empirical spontaneous traffic breakdown at the bottleneck.

- Qualitative characteristics of synchronized flow that has emerged at the bottleneck due to the breakdown do not depend on whether empirical spontaneous traffic breakdown or empirical induced traffic breakdown has occurred at the bottleneck.

Waves in Empirical Free Flow

In each of the freeway lanes (Fig. 11a, c), road detectors measure the following 1-min averaged data: the flow rate of all vehicles q , the average speed v , as well as the flow rate q_{slow} of long vehicles. Respectively, we can calculate the percentage of long vehicles $(q_{\text{slow}}/q)100\%$. Long vehicles can be considered *slow vehicles*. On the working days (Fig. 11b, d), the most of long



Modeling Approaches to Traffic Breakdown,

Fig. 11 Overview of features of empirical spontaneous traffic breakdown (F $\rightarrow$ S transition) at an on-ramp bottleneck (a, b) and an off-ramp bottleneck (c, d) (real field traffic data measured by road detectors on three-lane freeway A5-South in Germany on April 15, 1996 (b) and September 03, 1998 (d)): (a, c) Sketches of sections of three-lane highway in Germany with off- and on-ramps bottleneck. (b, d) Speed data measured with road detectors installed along road section in (a, c); data are presented in

space and time with averaging method described in section C.2 of (Kerner et al. 2013b). (b) Empirical spontaneous traffic breakdown at on-ramp bottleneck. (d) Empirical spontaneous traffic breakdown at off-ramp bottleneck. The on-ramp bottleneck labeled by “on-ramp bottleneck” in (a, b) is the same as that labeled by “on-ramp bottleneck 2” in (c, d). Off-ramp bottleneck, on-ramp bottleneck 1, and on-ramp bottleneck 2 marked by dashed lines in (c) are, respectively, the bottlenecks explained on Fig. 2.1 of the book (Kerner 2004a)

vehicles are trucks. Trucks have a speed limit 80 km/h on German highways (in reality, trucks move usually at the speed within a range 80–90 km/h).

To find nuclei for empirical spontaneous traffic breakdown, we study possible waves in free flow. Additionally with possible waves of 1-min average traffic variables q , q_{slow} , and v , we investigate also waves of the following variables

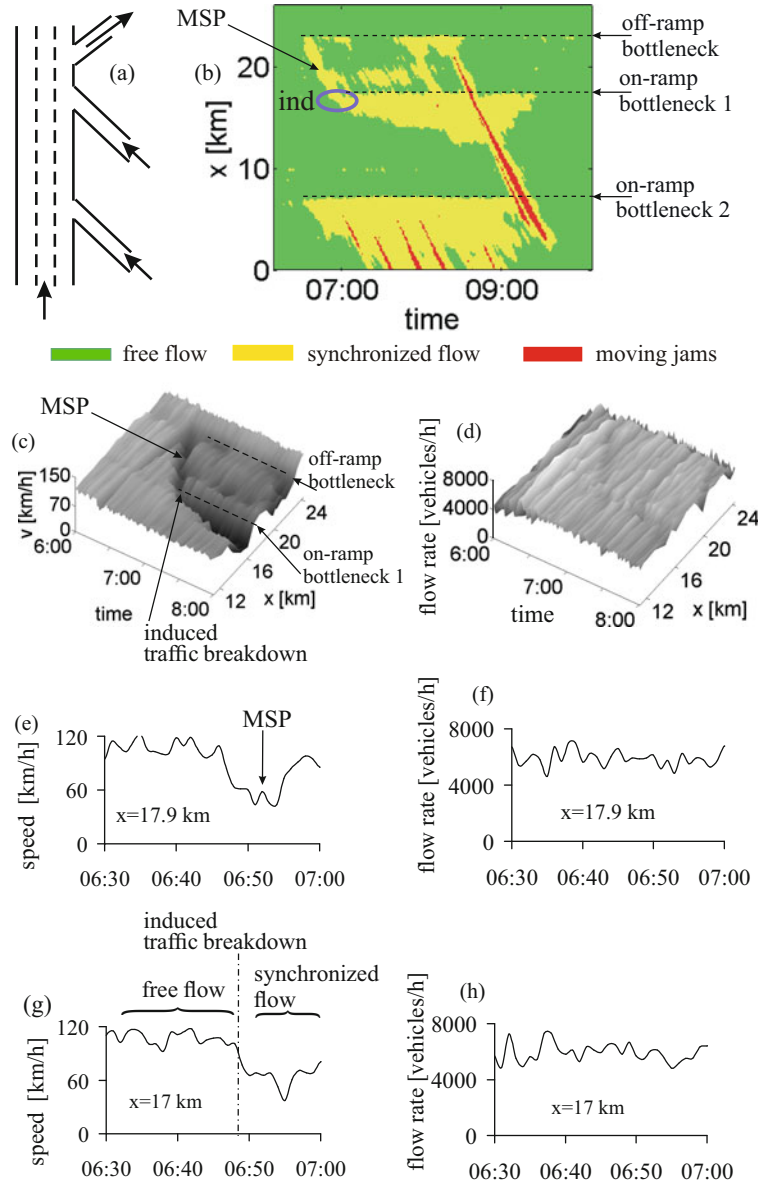
$$\Delta q_{\text{wave}} = q - \bar{q}, \quad \Delta v_{\text{wave}} = \bar{v} - v, \quad (3)$$

$$\Delta \psi_{\text{wave}} = \begin{cases} \Delta q_{\text{wave}}^{(\text{slow})} / \Delta q_{\text{wave}} & \text{at } \Delta q_{\text{wave}} > \delta_{\text{wave}}, \\ 0 & \text{at } \Delta q_{\text{wave}} \leq \delta_{\text{wave}}, \end{cases} \quad (4)$$

where $\Delta \psi_{\text{wave}}$ is a dimensionless characteristic of the share of long (slow) vehicles, δ_{wave} is constant ($\delta_{\text{wave}} > 0$); $\Delta q_{\text{wave}}^{(\text{slow})} = q_{\text{slow}} - \bar{q}_{\text{slow}}$; traffic

Modeling Approaches to Traffic Breakdown,

Fig. 12 Empirical example of traffic breakdown at on-ramp bottleneck induced by propagation of moving synchronized flow pattern (MSP): (a, b) Sketch of section of three-lane highway (a) and speed data (b); induced traffic breakdown is labeled by “ind” in (b). (c, d) Average speed (c) and total flow rate (d) on the main road in space and time related to a fragment of (b). (e–h) Time-functions of average speed (e, g) and total flow rate (f, h) at two road locations related to (c, d): 17.9 km (e, f) and 17 km (g, h). Location 17 km is about 100 m downstream of the end of the merging region of on-ramp bottleneck 1 (c). Real 1-min average data measured by detectors on three-lane freeway A5-South in Germany on April 20, 1998 (Adapted from Kerner 2004a)



variables $\bar{q}$, $\bar{q}_{\text{slow}}$, and $\bar{v}$ are related to 20-min average data with the use of the well-known procedure of “moving averaging.” The dimensionless characteristic of the share of long (slow) vehicles $\Delta\psi_{\text{wave}}$ allows us to make a wave of the percentage of long (slow) vehicles visible in space x and time t .

To make possible wave propagation in free flow visible in space and time, we apply an

approximate wave reconstruction procedure of (Kerner et al. 2015) (Fig. 13). We have found that empirical waves can propagate through the whole road section. Because vehicles enter the main road from on-ramps and leave the main road to off-ramps, some of the waves can appear at on-ramps or disappear at off-ramps.

We find out that the waves of the flow rate Δq_{wave} and the speed Δv_{wave} almost coincide

with the waves of the share of slow vehicles in free flow $\Delta\psi_{\text{wave}}$ (Fig. 13). We see that each of the waves propagates downstream with the mean wave velocity v_{wave} that is approximately equal to the mean speed of slow vehicles $v_{\text{slow}}^{(\text{mean})}$ (Fig. 13):

$$v_{\text{wave}} = v_{\text{slow}}^{(\text{mean})}. \quad (5)$$

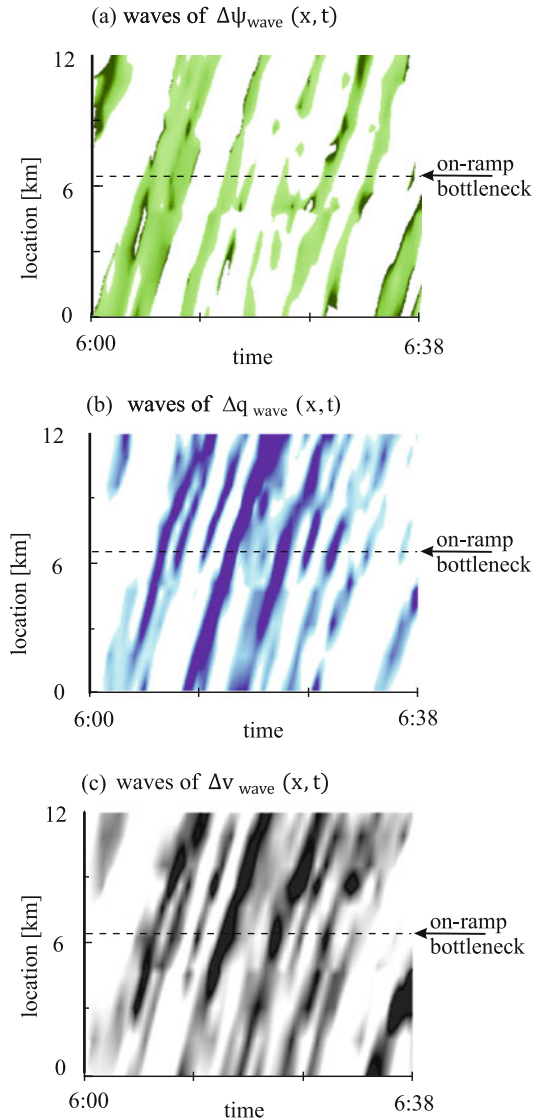
In all empirical data measured on working days, $v_{\text{slow}}^{(\text{mean})}$ is given by the average speed of long vehicles that changes within range 85–88 km/h. Within any of the waves propagating with the velocity v_{wave} (5), the percentage of long vehicles and the flow rate are larger, whereas the average speed is lower than outside the wave.

- In free flow, there are waves of the flow rate and the speed that coincide approximately with waves of the share of slow vehicles. The waves propagate in the flow direction on average with the average speed of slow vehicles. Within any of the waves, the share of slow vehicles and the flow rate are larger, whereas the average vehicle speed is lower than outside the wave.

Empirical Spontaneous Nucleation of Traffic Breakdown at On-Ramp Bottleneck

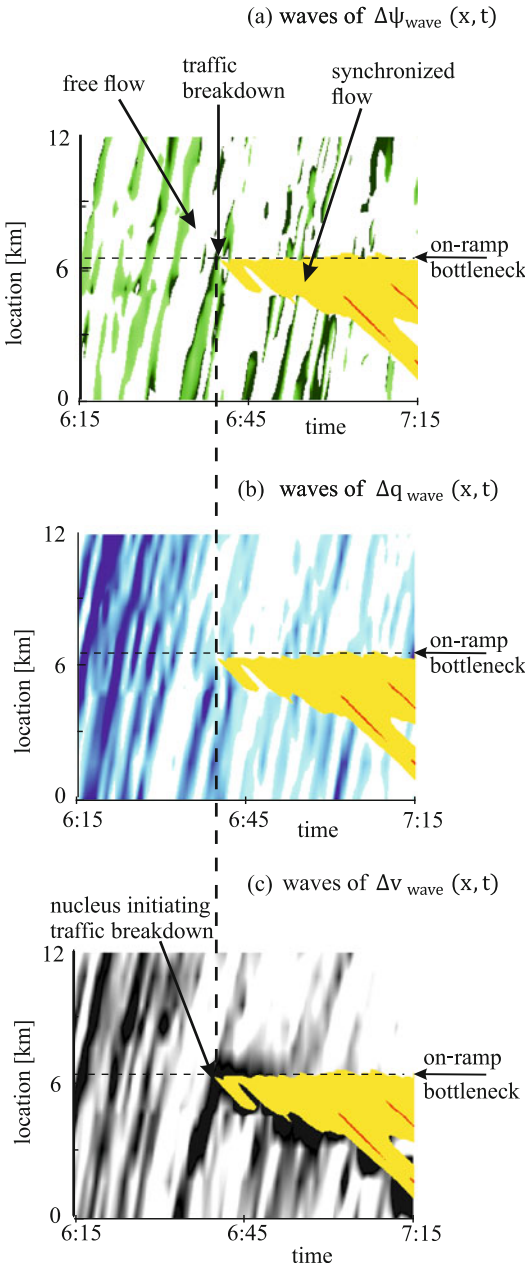
During a long time interval, the waves of traffic variables in free flow propagate with the positive velocity v_{wave} (5) through the on-ramp bottleneck. No traffic breakdown occurs at the bottleneck (Fig. 13). This changes crucially when we consider a longer time interval (Fig. 14).

Indeed, when one of the waves propagates through the on-ramp bottleneck, the wave initiates traffic breakdown at the bottleneck. During the subsequent wave propagation downstream of the bottleneck, the structure of the wave and its features do not change. Thus, one of the waves in free flow studied above leads to the occurrence of a nucleus for traffic breakdown at the bottleneck, when the wave propagates through the bottleneck (labeled by “nucleus initiating traffic breakdown” in Fig. 14c).



Modeling Approaches to Traffic Breakdown, Fig. 13

Empirical waves of $\Delta\psi_{\text{wave}}$ (a), Δq_{wave} (b), and Δv_{wave} (c) for data in Fig. 11b averaged across the road (real field traffic data measured by road detectors installed along three-lane freeway): (a) Waves of $\Delta\psi_{\text{wave}}(x, t)$ are presented by regions with variable shades of gray (green in the on-line version) (in white regions $\Delta\psi_{\text{wave}} \leq 0.1$, in black (dark green) regions $\Delta\psi_{\text{wave}} \geq 1$). (b) Waves of $\Delta q_{\text{wave}}(x, t)$ are presented by regions with variable shades of gray (blue in the on-line version) (in white regions $\Delta q_{\text{wave}} \leq 600$ vehicles/h, in black (dark blue) regions $\Delta q_{\text{wave}} \geq 1500$ vehicles/h). (c) Waves of $\Delta v_{\text{wave}}(x, t)$ are presented by regions with variable shades of gray (in white regions $\Delta v_{\text{wave}} \leq 1$ km/h, in black regions $\Delta v_{\text{wave}} \geq 15$ km/h).



Modeling Approaches to Traffic Breakdown, Fig. 14 Empirical nucleus in free flow for data in Fig. 11b. Empirical waves of $\Delta\psi_{\text{wave}}$ (a), Δq_{wave} (b), and Δv_{wave} (c) in free flow for a longer time interval as that in Fig. 13. In (a–c), regions labeled by “synchronized flow” show symbolically synchronized flow. Parameters of the presentation of empirical waves in (a–c) are the same as those in Fig. 13a–c, respectively

In chapter “► [Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks](#)” of the book (Kerner 2017a), it has been shown that a wave becomes a nucleus for traffic breakdown *only* at the effective location of the bottleneck at which a permanent speed disturbance is localized.

- The physics of the occurrence of empirical nuclei for traffic breakdown at highway bottlenecks can be explained by an interaction of a wave in free flow with a permanent speed disturbance localized at the effective location of the bottleneck.

Induced Traffic Breakdown – Empirical Proof of Nucleation Nature of Empirical Traffic Breakdown

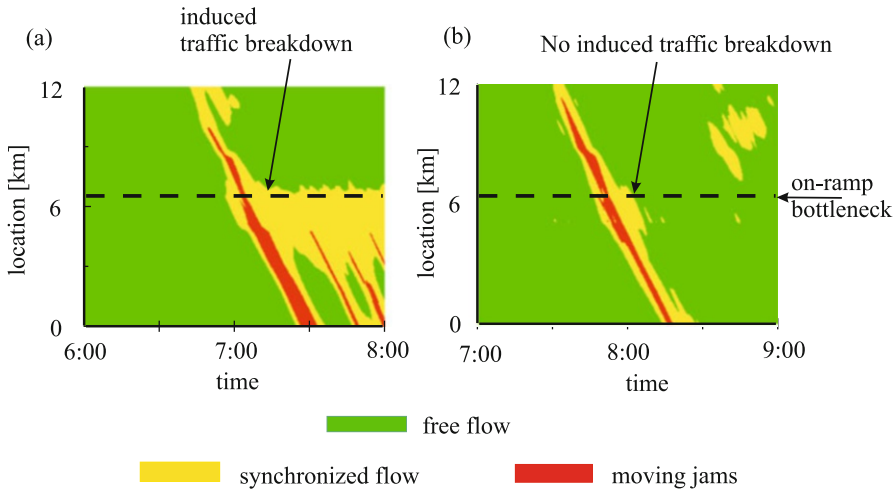
The author has often confronted with the following question of colleagues:

- Is there some general empirical proof of the nucleation nature of empirical traffic breakdown ($F \rightarrow S$ transition)?

To perform a clear empirical proof of the nucleation nature of traffic breakdown that is independent of differences in vehicle and driver characteristics in free flow, we compare *empirical induced* traffic breakdown (Fig. 15a) with *empirical spontaneous* traffic breakdown caused by wave propagation (Fig. 14).

Both an empirical wave in free flow (Fig. 14) and a localized congested pattern (wide moving jam in Fig. 15a) become nuclei for traffic breakdown in metastable free flow, when they reach the effective location of a highway bottleneck. The propagation of a single congested pattern to the effective bottleneck location is usually sufficient for the inducing of the breakdown at the bottleneck (Fig. 15a). In contrast, many waves in free flow can propagate through the bottleneck while initiating no breakdown at the bottleneck (Fig. 13).

The latter empirical result allows us to assume that at a given flow rate in metastable free flow at a highway bottleneck only a wave that amplitude is large enough can cause the occurrence of a *critical nucleus* required for traffic breakdown while the wave propagates through the bottleneck.



Modeling Approaches to Traffic Breakdown, Fig. 15 Empirical jam propagation through on-ramp bottleneck: (a) Empirical induced traffic breakdown. (b) Moving jam propagation through the bottleneck without induced traffic breakdown. Real field traffic data measured by road detectors on three-lane freeway A5-South in

Germany on March 22, 2001 (a) and June 23, 1998 (b). On-ramp bottleneck marked by dashed lines in (a, b) is the same one as that in Fig. 11a, b. Speed data measured with road detectors installed along road section; data is presented in space and time with averaging method described in section C.2 of (Kerner et al. 2013b)

Therefore, if the amplitude of a wave is small enough, then no nucleus for traffic breakdown occurs when the wave propagates through the bottleneck; respectively, no traffic breakdown is observed at the bottleneck (Fig. 13).

In contrast with waves in free flow, within a congested pattern the speed is usually smaller than the speed within a critical nucleus required for the breakdown in metastable free flow at a bottleneck. For this reason, a localized congested pattern becomes a nucleus for traffic breakdown in the metastable free flow, when the pattern reaches the effective bottleneck location.

Thus, a basic difference between *empirical spontaneous* breakdown (Figs. 11b, d and 14) and *empirical induced* breakdown (Fig. 15a) is as follows: To initiate spontaneous traffic breakdown at the bottleneck in metastable free flow, the amplitude of a wave that causes the occurrence of a nucleus for traffic breakdown should be large enough. In contrast, a localized congested pattern is usually a nucleus for the breakdown in the metastable free flow at the bottleneck.

However, after the breakdown has occurred, characteristics of synchronized flow that has been formed at the bottleneck do not depend on whether synchronized flow has occurred due to

empirical spontaneous breakdown or due to empirical induced breakdown. This statement has been empirically proven in Chapter “► [Dynamic Traffic Routing, Assignment, and Assessment of Traffic Networks](#)” of the book (Kerner 2017a).

This shows that rather than the nature of traffic breakdown, the terms *empirical spontaneous* and *empirical induced* traffic breakdowns at a bottleneck distinguish different *sources* of a nucleus that occurrence leads to traffic breakdown: In Fig. 11b, the source of empirical spontaneous traffic breakdown is one of the waves in free flow shown in Fig. 14. In Fig. 15a, the source of empirical induced traffic breakdown is the wide moving jam. In Fig. 12, the source of empirical induced traffic breakdown is the MSP.

The empirical induced traffic breakdown at a highway bottleneck discussed above does not depend on the degree of the heterogeneity of real vehicular traffic. Empirical induced traffic breakdown is a general empirical proof of the nucleation nature of empirical traffic breakdown.

- Rather than the nature of traffic breakdown, the terms *empirical spontaneous* and *empirical induced* traffic breakdowns at a bottleneck distinguish different *sources* of a nucleus that

occurrence leads to traffic breakdown at the bottleneck.

- The empirical induced traffic breakdown at a highway bottleneck is a general empirical proof of the nucleation nature of empirical traffic breakdown: This empirical proof is independent of the degree of the heterogeneity of real vehicular traffic.

Minimum Highway Capacity as Threshold for Empirical Induced Traffic Breakdown

In contrast with the wide moving jam shown in Fig. 15a, a wide moving jam shown in Fig. 15b does not induce traffic breakdown at the bottleneck: After the wide moving jam shown Fig. 15b is far away upstream of the bottleneck, free flow returns both at the effective bottleneck location and downstream and upstream of the bottleneck.

However, if free flow is in a metastable state at the bottleneck, the wide moving jam should be a nucleus for traffic breakdown at the bottleneck. Therefore, the case shown in Fig. 15b, in which no traffic breakdown has been induced, should be related to the opposite condition: Rather than free flow is metastable, this free flow is in a stable state with respect to traffic breakdown ($F \rightarrow S$ transition).

We can conclude that there should be some critical flow rate that separates metastable and stable states of free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. When the flow rate at the bottleneck is smaller than this critical flow rate, there can be no nuclei for traffic breakdown at the bottleneck. This means that regardless how large a time-limited disturbance in free flow at the bottleneck is, no traffic breakdown occurs at the bottleneck: After the disturbance has disappeared at the bottleneck, free flow recovers at the bottleneck. This critical flow rate determines the minimum highway capacity $C_{\min}$ briefly discussed in section “The Basic Assumption of Three-Phase Traffic Theory.” This emphasizes another difference between empirical spontaneous and empirical induced traffic breakdowns at a highway bottleneck that is as follows.

As we have seen above, there are many waves in free flow that propagate through the bottleneck

without initiating of empirical spontaneous traffic breakdown. From this empirical result, we *cannot* state whether free flow is in a metastable state or free flow is in a stable state with respect to traffic breakdown. Indeed, if all these waves are smaller than a critical wave, then no traffic breakdown occurs in the metastable free flow at the bottleneck. No traffic breakdown occurs also, when free flow is in a stable state with respect to traffic breakdown, i.e., if the flow rate in free flow at the bottleneck is smaller than the minimum highway capacity $C_{\min}$.

In contrast, when a wide moving jam propagates through the bottleneck without inducing of traffic breakdown, we can assume that free flow is in a stable state with respect to traffic breakdown ($F \rightarrow S$ transition). This means that when the wide moving jam propagates through a bottleneck without inducing the breakdown, we can assume that the flow rate is smaller than the minimum highway capacity $C_{\min}$ of free flow at the bottleneck.

The Basic Assumption of Three-Phase Traffic Theory

To explain features of empirical spontaneous and empirical induced traffic breakdowns at highway bottlenecks (section “Empirical Proof of Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition)”), in the three-phase theory the following basic assumption has been made (Kerner 1998a, b, 1999a, b, 2004a)

$$P^{(B)}(q_{\text{sum}}) = P_{\text{nucleus}}^{(B)}(q_{\text{sum}}). \quad (6)$$

In Eq. 6, $P^{(B)}(q_{\text{sum}})$ is the flow-rate dependence of the probability that during a given time interval T_{ob} for observing traffic flow spontaneous traffic breakdown occurs at a highway bottleneck, $P_{\text{nucleus}}^{(B)}(q_{\text{sum}})$ is the flow-rate dependence of the probability that during the time interval T_{ob} a nucleus for traffic breakdown occurs spontaneously in free flow at the bottleneck.

A theoretical probability of spontaneous traffic breakdown in the framework of the microscopic

stochastic three-phase theory was firstly found in 2002 (Kerner et al. 2002). It has been found that the flow-rate function of the breakdown probability is well fitted by a function (Kerner et al. 2002) (dashed curves in Fig. 8):

$$P^{(B)}(q_{\text{sum}}) = \frac{1}{1 + \exp[\beta(q_p - q_{\text{sum}})]}, \quad (7)$$

where q_{sum} is the flow rate at the bottleneck, β and q_p are parameters.

It should be noted that a mathematical nucleation theory of traffic breakdown at a highway bottleneck (Kerner and Klenov 2005, 2006a, b) can be found in the article (Kerner and Klenov 2018) of this Encyclopedia.

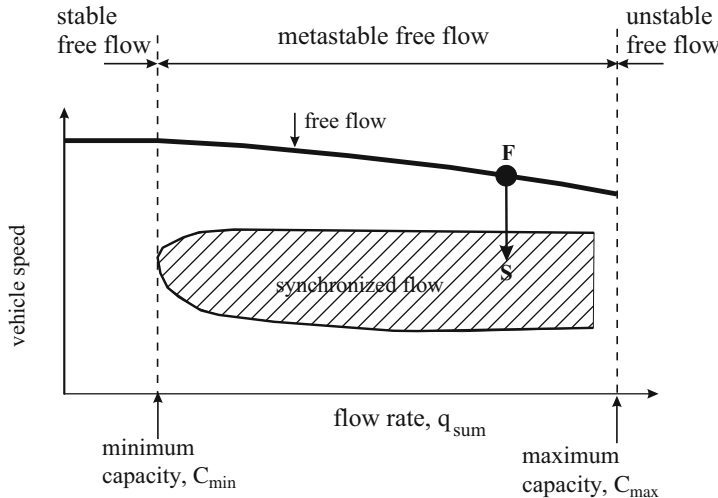
Empirical Nucleation Nature of Traffic Breakdown as Origin of the Infinity of Highway Capacities at Any Time Instant

We have explained that the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) observed above in real field traffic data means that free flow is in a metastable state with respect

to the $F \rightarrow S$ transition at the bottleneck. In its turn, the term “the metastability of free flow” means that if a nucleus for traffic breakdown occurs in a metastable state of free flow at the bottleneck, traffic breakdown does occur. In contrast, as long as no nucleus appears, no traffic breakdown occurs in the metastable state of free flow.

Clearly, there should be a *limited range* of the flow rate q_{sum} in free flow at the bottleneck within which empirical free flow is in a metastable state (Fig. 16). Indeed, in real field traffic data at small enough values of the flow rate q_{sum} , no traffic breakdown is observed at the bottleneck. This means that no nuclei can occur in free flow that can lead to traffic breakdown at the bottleneck. This free flow is stable with respect to traffic breakdown at the bottleneck. As explained above, the minimum flow rate that separates the stable and metastable free flow is equal to a minimum highway capacity $C_{\min}$ of free flow at the bottleneck (Fig. 16). Therefore, at

$$q_{\text{sum}} < C_{\min}, \quad (8)$$



Modeling Approaches to Traffic Breakdown, Fig. 16 Qualitative explanation of the empirical metastability of free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at bottleneck (Kerner 1998a, b, 1999a, 2004a). Qualitative Z-speed–flowrate characteristic for traffic breakdown (results of simulations of the Z-speed–flow-

rate characteristic for traffic breakdown will be discussed in section “[Theoretical Z-Characteristic for Traffic Breakdown at Bottleneck](#)”; see Fig. 36). Arrow $F \rightarrow S$ illustrates symbolically one of possible $F \rightarrow S$ transitions occurring in a metastable state of free flow when a nucleus appears in the free flow at the bottleneck

free flow is stable at the bottleneck. In other words, under condition (8) no nuclei can appear in the free flow. Consequently, no traffic breakdown ($F \rightarrow S$ transition) can occur at the bottleneck.

Contrarily, at

$$q_{\text{sum}} \geq C_{\min} \quad (9)$$

free flow is in a metastable state with respect to traffic breakdown at the bottleneck (Fig. 16). Therefore, under condition (9), traffic breakdown can occur at the bottleneck.

There should be a large enough flow rate at which any small disturbance in free flow is a nucleus for traffic breakdown at the bottleneck. At this flow rate that we denote by a maximum highway capacity $C_{\max}$ of free flow at the bottleneck free flow can be considered as unstable free flow (Fig. 16). This means that the flow rate

$$q_{\text{sum}} = C_{\max} \quad (10)$$

separates the metastable free flow and unstable free flow with respect to traffic breakdown at the bottleneck (Fig. 16): At

$$q_{\text{sum}} \geq C_{\max} \quad (11)$$

free flow is unstable with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck.

Correspondingly, when both condition (9) and condition

$$q_{\text{sum}} < C_{\max} \quad (12)$$

are satisfied, free flow is metastable with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. Thus, from formulas (9) and (12), it follows that free flow is in a metastable state with respect to the $F \rightarrow S$ transition at the bottleneck within the flow rate range (Fig. 16)

$$C_{\min} \leq q_{\text{sum}} < C_{\max}. \quad (13)$$

Any flow rate q_{sum} in free flow at a highway bottleneck that satisfies (13) is highway capacity. This is because at each of the flow rates (13) traffic

breakdown can occur at the bottleneck. At any time instant, there are the infinite number of the flow rates (13) at which traffic breakdown can occur at the bottleneck. Therefore, there are an infinite number of highway capacities at any time instant. The existence of an infinite number of highway capacities at any time instant means that highway capacity is stochastic. This conclusion is one of the important consequences of the empirical nucleation nature of traffic breakdown discussed in this chapter.

We can make the following conclusion:

- *At any time instant*, there are the infinite number of highway capacities C of free flow at the bottleneck. The range of the infinite number of highway capacities is limited by the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$ (Fig. 16):

$$C_{\min} \leq C \leq C_{\max}, \quad (14)$$

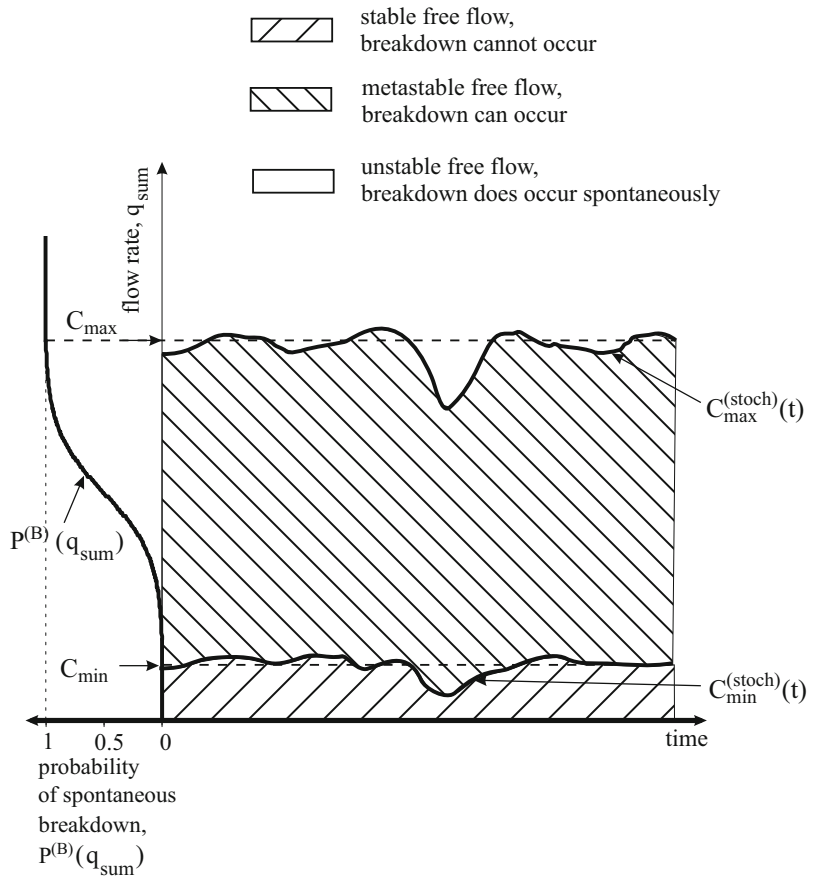
where $C_{\min} < C_{\max}$. The existence of an infinite number of highway capacities at any time instant means that highway capacity is stochastic.

It must be noted that the maximum capacity $C_{\min}$ and the minimum capacity $C_{\max}$ considered above depend on traffic parameters. These traffic parameters are weather, mean driver's characteristics (e.g., mean driver reaction time), share of long vehicles, etc. In real traffic flow, these traffic parameters change over time. For this reason, the values of the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$ change also over time. Moreover, in real traffic flow, the traffic parameters are stochastic time functions. Therefore, in real traffic flow, we should consider some stochastic maximum capacity $C_{\max}^{(\text{stoch})}(t)$ and stochastic minimum capacity $C_{\min}^{(\text{stoch})}(t)$ whose time dependencies are determined by stochastic characteristics of traffic parameters. Qualitative hypothetical fragment of these time-functions within a time interval is shown in Fig. 17.

Thus, we should generalize the above definition of stochastic highway capacity as follows:

Modeling Approaches to Traffic Breakdown,

Fig. 17 Qualitative explanations of a connection between the probability of traffic breakdown at bottleneck and stochastic highway capacity



- At any time instant t , there are the infinite number of stochastic highway capacities C of free flow at a highway bottleneck. The range of these capacities is limited by a minimum stochastic highway capacity $C_{\text{min}}^{(\text{stoch})}(t)$ and a maximum stochastic highway capacity $C_{\text{max}}^{(\text{stoch})}(t)$ (Fig. 17):

$$C_{\text{min}}^{(\text{stoch})}(t) \leq C \leq C_{\text{max}}^{(\text{stoch})}(t), \quad (15)$$

where $C_{\text{min}}^{(\text{stoch})}(t) < C_{\text{max}}^{(\text{stoch})}(t)$.

It should be noted that stochastic functions $C_{\text{max}}^{(\text{stoch})}(t)$ and $C_{\text{min}}^{(\text{stoch})}(t)$ (Fig. 17) cannot be measured in empirical observations. Only their mean values, respectively, C_{max} and C_{min} can be found in empirical studies of measured traffic data.

If at a given time instant t the flow rate $q_{\text{sum}}(t)$ is related to a range of the flow rate in Fig. 17 that satisfies condition

$$q_{\text{sum}}(t) < C_{\text{min}}^{(\text{stoch})}(t), \quad (16)$$

no traffic breakdown can occur at the bottleneck. Thus, this flow rate is smaller than highway capacity. In this stable free flow (16), no traffic breakdown (F $\rightarrow$ S transition) can occur at the bottleneck.

If at a given time instant t the flow rate $q_{\text{sum}}(t)$ is related to a range of the flow rate in Fig. 17 that satisfies conditions

$$C_{\text{min}}^{(\text{stoch})}(t) \leq q_{\text{sum}}(t) < C_{\text{max}}^{(\text{stoch})}(t), \quad (17)$$

free flow is in a metastable state with respect to traffic breakdown at the bottleneck. In other

words, this means that in this case traffic breakdown ($F \rightarrow S$ transition) can occur at the bottleneck.

- The term “stochastic” in the definition of stochastic highway capacity made in the framework of the three-phase theory is explained by the possibility of *random occurrence* of a nucleus for traffic breakdown. The nucleus can occur at any flow rate $q_{\text{sum}}(t)$ in free flow at the bottleneck that satisfies conditions (17).

If at a given time instant t the flow rate $q_{\text{sum}}(t)$ is related to a range of the flow rate in Fig. 17 that satisfies conditions

$$q_{\text{sum}}(t) \geq C_{\text{max}}^{(\text{stoch})}(t), \quad (18)$$

free flow is in an unstable state with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. In other words, in this case traffic breakdown does occur spontaneously during the time interval T_{ob} at the bottleneck.

Failure of Classical Understanding of Stochastic Highway Capacity

In the classical understanding of stochastic highway capacity (see section “[Classical Understanding of Stochastic Highway Capacity](#)”), free flow is stable under condition

$$q_{\text{sum}}(t) < C(t). \quad (19)$$

This means that no traffic breakdown can occur or be induced at the bottleneck as long as the flow rate in free flow at the bottleneck is smaller than the stochastic capacity. This contradicts to the empirical fact that traffic breakdown can be induced at the bottleneck due to the upstream propagation of a localized congested pattern.

This is because stochastic highway capacity cannot depend on whether there is a congested pattern, which has occurred outside the bottleneck and independent of the bottleneck existence, or not. Indeed, the empirical evidence of induced

traffic breakdown is the empirical proof that at a given flow rate at a bottleneck, there can be one of two different traffic states at the bottleneck:

- (i) A traffic state related to free flow and
- (ii) A congested traffic state related to synchronized flow

Due to the upstream propagation of a localized congested pattern, a transition from the state of free flow to the state of synchronized flow, i.e., traffic breakdown is induced.

The induced traffic breakdown is impossible to occur under the classical understanding of the nature of highway capacity (Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990). This is because in the classical understanding of highway capacity, free flow is stable under condition (19), i.e., no traffic breakdown can occur (Fig. 10a).

In contrast with the classical understanding of the nature of stochastic highway capacity, the evidence of the empirical induced breakdown means that free flow is in a metastable state with respect to the breakdown. The metastability of free flow at the bottleneck should exist for all flow rates at which traffic breakdown can be induced at the bottleneck. This empirical evidence of the metastability of free flow at the bottleneck contradicts fundamentally the concept of Brilon’s stochastic capacity, in which free flow is stable under condition (19).

Thus, the currently accepted understanding of stochastic highway capacity (Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990) failed because this understanding about the nature of highway capacity contradicts the empirical evidence that traffic breakdown can be induced at a highway bottleneck as observed in real traffic.

The observation of empirical induced breakdowns proves that condition (2) of Brilon’s stochastic highway capacity (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014;

Elefteriadou et al. 2014) cannot be valid for real traffic. However, the following question arises:

- What are the consequences of this controversial understanding of the nature of traffic breakdown in the classical theory (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 2014) and in the three-phase theory (Kerner 2004a, 2017a)?

The classical understanding of stochastic highway capacity is based on the assumption that the empirical probability of traffic breakdown is determined by the capacity distribution function, i.e., condition (2) is valid. In contrast, the assumption of the three-phase theory about the metastability of traffic breakdown with respect to traffic breakdown (6) is based on the evidence that empirical traffic breakdown can be induced at a bottleneck. In both the classical theory and three-phase traffic theory, there is the infinite number of stochastic highway capacities. However, in the classical understanding of stochastic highway capacity at a *given time instant*, there is *only one* value of highway capacity (Fig. 9).

Contrarily, in the three-phase theory at a *given time instant*, there are *the infinite number* of highway capacities (Fig. 17). The existence at any time instant of the infinite number of highway capacities means that highway capacity is stochastic. A infinite number of stochastic capacities are within some capacity range between minimum $C_{\min}^{(\text{stoch})}$ and maximum highway capacities $C_{\max}^{(\text{stoch})}$. We do not know values of $C_{\min}^{(\text{stoch})}$ and $C_{\max}^{(\text{stoch})}$ because $C_{\max}^{(\text{stoch})}(t)$ and $C_{\min}^{(\text{stoch})}(t)$ are stochastic values. Due to the evidence of the possibility of empirical induced traffic breakdown, we know only that $C_{\max}^{(\text{stoch})}(t) > C_{\min}^{(\text{stoch})}(t)$ (Fig. 17).

With the use of Fig. 18, we can qualitatively illustrate the basic difference between the classical understanding of stochastic highway capacity and the understanding of the infinite number of stochastic highway capacities made in the three-phase theory. In the classical understanding of stochastic capacity (2), for the hypothetical time

dependence of the flow rate $q_{\text{sum}, 2}(t)$ shown in Fig. 10b, traffic breakdown has occurred at time instant t_1 at which condition $q_{\text{sum}, 2}(t) = C(t)$ is satisfied, i.e., when the flow rate is equal to the capacity value. In contrast, in accordance with the three-phase theory for the same time dependence of the flow rate $q_{\text{sum}, 2}(t)$, for which conditions (17) are satisfied, no breakdown should be necessarily occur both at time instant t_1 and for a later time interval (Fig. 18a).

In the classical understanding of stochastic capacity (2), for the hypothetical time dependence of the flow rate $q_{\text{sum}, 1}(t)$ shown in Fig. 10a, traffic breakdown could not occur because for all time instants condition $q_{\text{sum}}(t) < C(t)$ (19) is satisfied. In contrast, in accordance with the three-phase theory for the same time dependence of the flow rate $q_{\text{sum}, 1}(t)$, when the flow rate $q_{\text{sum}, 1}(t)$ satisfies condition (17), free flow is in a metastable state with respect to $F \rightarrow S$ transition at the bottleneck; this means that traffic breakdown can occur spontaneously in this metastable free flow as this is shown for time instant t_2 in Fig. 18b.

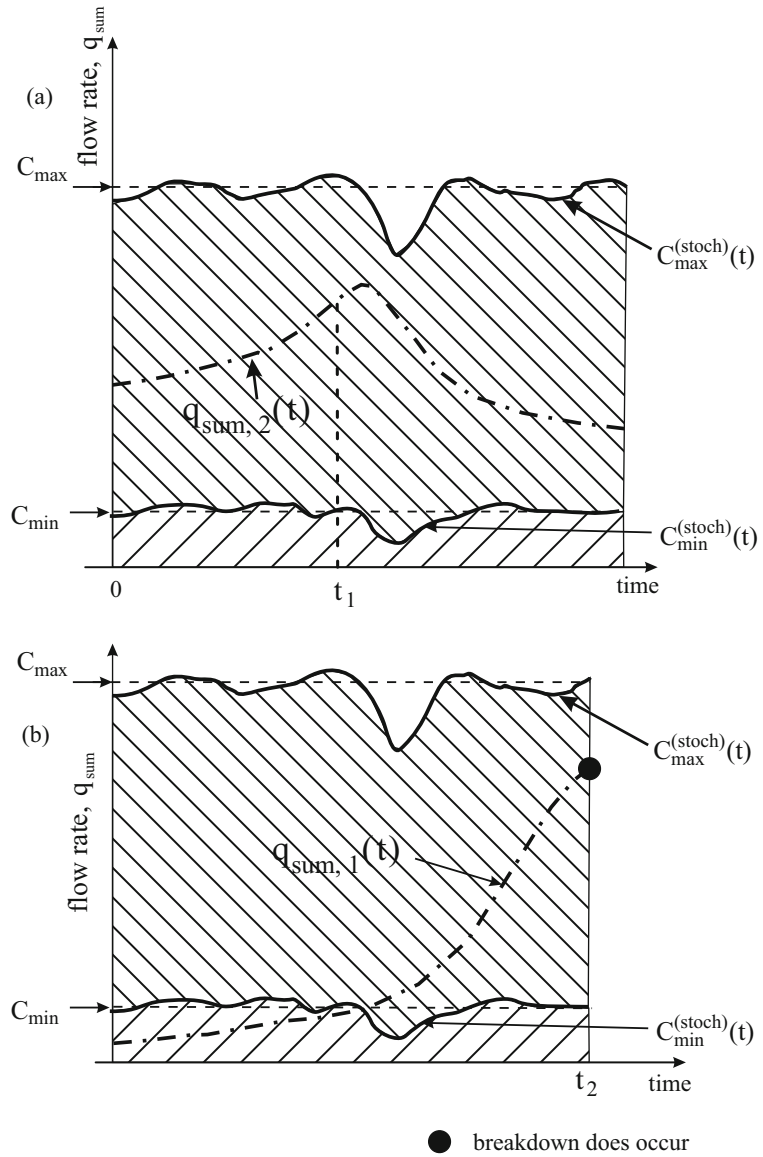
Because the classical understanding of stochastic highway capacity (1), (2) contradicts the empirical nucleation nature of real traffic breakdown, the understanding of stochastic highway capacity made in (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 2014) cannot be used for reliable highway design and highway operations. This critical conclusion is also related to a diverse variety of ITS-applications of the classical understanding of highway capacity (see an example in section “ANCONA On-Ramp Metering”).

Failure of Classical Modeling Approaches to Traffic Breakdown at Highway Bottlenecks

Traffic researchers have developed a huge number of traffic theories for optimization and control of traffic and transportation networks. In particular, to generally accepted fundamentals and methodologies of traffic and transportation theory belong the following theories and associated methodologies:

Modeling Approaches to Traffic Breakdown,

Fig. 18 Qualitative explanation of traffic breakdown in the three-phase theory in which at any time instant there is the infinite number of capacities of free flow at a highway bottleneck. Hypothetical time-functions $C_{\max}^{(\text{stoch})}(t)$ and $C_{\min}^{(\text{stoch})}(t)$ are taken from Fig. 17. Hypothetical time functions of the flow rates $q_{\text{sum}, 2}(t)$ in (a) and $q_{\text{sum}, 1}(t)$ in (b) as well as time instant t_1 in (a) are, respectively, the same as those in Fig. 10 (Adapted from Kerner 2017a)



- (i) The Lighthill-Whitham-Richards (LWR) model (Lighthill and Whitham 1955; Richards 1956).
- (ii) The General Motors (GM) model class.

In traffic flow models within the framework of the GM model class traffic breakdown is explained by a free flow instability that causes a growing wave of vehicle speed reduction propagating upstream in traffic flow. This classical traffic flow instability has been introduced in the GM car-following

model by Herman, Gazis, Rothery, Montroll, Chandler, and Potts introduced in 1959–1961 (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959) (see, e.g., (Bellomo et al. 2002; Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Treiber and Kesting 2013; Whitham 1974; Wiedemann

- 1974; Wolf 1999) as well as references in reviews (Kerner 2015b; Kerner 2016) and the book (Kerner 2017a)).
- (iii) The understanding of highway capacity of free flow at a bottleneck as a stochastic value (see, e.g., Banks 1989, 1990; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990).
 - (iv) Principles for standard dynamic traffic assignment in a traffic and transportation network that should minimize travel time (or other travel costs) in the network (examples are Wardrop's user equilibrium (UE) and system optimum (SO) principles (Wardrop 1952)).

These classical theories and methodologies are the basis for dynamic traffic assignment, traffic control, and optimization in traffic and transportation networks as well as for development and simulations of intelligent transportation systems (ITS), traffic simulation tools, and for many other traffic engineering applications (see, e.g., (Ahn and Cassidy 2007; Banks 1989, 1990; Bellomo et al. 2002; Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Chowdhury et al. 2000; Cremer 1979; Daganzo 1994; Daganzo 1995, 1997; Elefteriadou 2014; Elefteriadou et al. 1995, 2014; Gartner et al. 1997; Gazis 2002; Haight 1963; Hall and Agyemang-Duah 1991; Hall et al. 1992; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Papageorgiou et al. 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999; Zurlinden 2003) as well as references in chapters “► [Complex Dynamics of Traffic Management: Introduction](#)” and “► [Optimization and Control of Urban Traffic Networks](#)” of the book (Kerner 2017a)).

The fundamentals and associated methodologies of these classical traffic theories have made a great impact on the understanding of many traffic

phenomena. However, approaches for control, dynamic traffic assignment, and network optimization based on these fundamentals and methodologies have failed by their applications in the real world. Even several decades of a very intensive effort to improve and validate network optimization models have no success. Indeed, there can be found no examples where on-line implementations of the network optimization models based on these fundamentals and methodologies could reduce congestion in real traffic and transportation networks.

In the review article (Kerner 2017b) and in chapter “► [Travel Behavior and Demand Analysis and Prediction](#)” of the book (Kerner 2017a), we have explained that and why any application of standard dynamic traffic assignment deteriorates basically the traffic system in a network while provoking heavy traffic congestion in the network. In section “[Failure of Classical Understanding of Stochastic Highway Capacity](#)” of this article, we have already explained why the classical understanding of stochastic highway capacity (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 2014) contradicts basically to the empirical fundamental of transportation science – the empirical nucleation nature of traffic breakdown.

The objective of this section is to explain the failure of applications of the generally accepted classical traffic flow models for the control and optimization of traffic and transportation networks in the real world. We show that there is *a fundamental problem* for these classical traffic flow models that is as follows:

1. Generally accepted fundamentals and methodologies of classical traffic and transportation theories are not consistent with the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck.
2. For this reason, traffic flow models based on the fundamentals and methodologies of traffic and transportation theory cannot be applied for reliable control, dynamic traffic assignment, and optimization in traffic and transportation networks as well as for other ITS-applications.

Classical Lighthill-Whitham-Richards (LWR) Theory of Onset of Traffic Congestion

The basic idea of the classical LWR traffic flow theory is as follows: the maximum flow rate q_0 associated with the maximum point (ρ_0, q_0) at the fundamental diagram (Fig. 1a) determines free flow capacity at a bottleneck (often called “bottleneck capacity”) denoted by q_{cap} :

$$q_{\text{cap}} = q_0. \quad (20)$$

It is further assumed that due to a road non-homogeneity introduced by the bottleneck the maximum flow rate q_0 downstream of the bottleneck, which determines the capacity (20), is smaller than the capacity of a homogeneous road upstream of the bottleneck. For this reason, if the sum of the flow rates upstream of the bottleneck reaches the maximum flow rate q_0 at the fundamental diagram (Fig. 1a), i.e., it reaches the bottleneck capacity (20), then a further increase in the flow rates must lead to congestion (queue) formation and upstream congestion propagation upstream of the bottleneck.

These hypotheses of the LWR theory mean that congested traffic occurs only then, when the upstream flow rate exceeds the bottleneck capacity q_{cap} determined by (20). This conclusion of the LWR traffic flow theory about the reason for the onset of congestion is very often used as the definition of congested traffic. As we will see below, this hypothesis of the LWR theory about the onset of congestion as well as the associated

congested traffic definition is in a very deep contradiction with empirical results.

In accordance with the LWR theory, traffic flow phenomena should be explained based on the law of conservation of the number of vehicles on the road

$$\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial Q(\rho(x, t))}{\partial x} = 0, \quad (21)$$

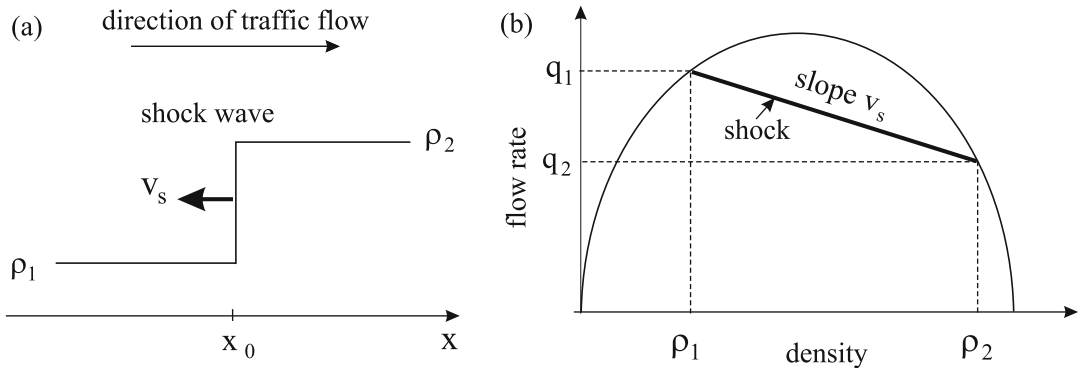
in which it is assumed that there is a relationship between the flow rate Q and density ρ

$$Q = Q(\rho) \quad (22)$$

given by the fundamental diagram for traffic flow. Here, x is a spatial coordinate in the direction of traffic flow and t is time. The LWR-model (21), (22) has discontinuous solutions in the form of shock-waves (Fig. 19) with a shock-wave velocity

$$v_s = \frac{Q(\rho_2) - Q(\rho_1)}{\rho_2 - \rho_1} \quad (23)$$

where the flow rates $Q(\rho_2)$, $Q(\rho_1)$ associated with (22) correspond to points in the flow–density plane that lie on the fundamental diagram; ρ_1 and ρ_2 are the vehicle densities upstream and downstream of the shock wave, respectively (Fig. 19a). The shock wave can be represented in the flow–density plane by the line labeled “shock” whose slope is equal to the shock velocity v_s (Fig. 19b).



Modeling Approaches to Traffic Breakdown,
Fig. 19 Qualitative representation of a shock wave in space (a) and in the fundamental diagram with the line

labeled “shock” whose slope is equal to the shock velocity v_s (b) (Adapted from Lighthill and Whitham 1955)

- In accordance with the LWR traffic flow theory, highway capacity is determined by the maximum flow rate q_0 at the fundamental diagram of traffic flow.

Daganzo's Cell Transmission Model

Because the LWR-model (21), (22) has discontinuous solutions in the form of shock waves, its numerical simulations are usually based on one of finite difference approximation methods for partial differential equations associated with the Godunov family methods. One of such discrete versions of the LWR-model (21) that is consistent with the LWR-hydrodynamic theory is the cell-transmission model of Daganzo (1994, 1995).

In Daganzo's cell transmission model, a rectangular lattice with time spacing τ and space length (cell length) Δx is overlaid on the time-space plane; the x-coordinates represent the center of the cells into which a road has been discretized; t-coordinates are the times at which the cell vehicle densities are evaluated. Then, the Daganzo model (Daganzo 1994, 1995) for a single-lane road of length L with an on-ramp bottleneck can be written as follows:

$$\begin{aligned} \rho(x, t + \tau) = & \rho(x, t) \\ & - [q(x + \Delta x/2, t + \tau/2) - q(x - \Delta x/2, t + \tau/2)] \\ & \times (\tau/\Delta x) + q_{\text{on}}(t)f(x)(\tau/\Delta x), \end{aligned} \quad (24)$$

$$\begin{aligned} q(x + \Delta x/2, t + \tau/2) = & \min(D(\rho(x, t)), R(\rho(x + \Delta x, t))), \end{aligned} \quad (25)$$

$$f(x) = \begin{cases} 1/N_m & \text{if } x_{\text{on}} \leq x < x_{\text{on}} + L_m, \\ 0 & \text{otherwise,} \end{cases} \quad (26)$$

with boundary conditions

$$\rho(0, t) = \rho(\Delta x, t), \rho(L, t) = \rho(L - \Delta x, t), \quad (27)$$

and initial conditions

$$\rho(x, 0) = \rho_{\text{in}}, \quad (28)$$

where $R(\rho)$ and $D(\rho)$ are nonincreasing so-called *receiving* and nondecreasing so-called *sending* curves, respectively; these two monotonic curves

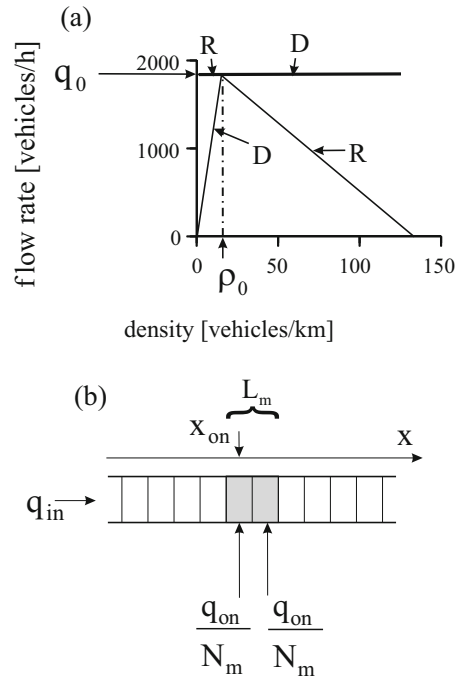
$R(\rho)$ and $D(\rho)$ take values in the interval $[0, q_0]$, as shown in Fig. 20a:

$$R(\rho) = \begin{cases} Q(\rho) & \text{if } \rho \geq \rho_0, \\ Q(\rho_0) & \text{if } \rho < \rho_0, \end{cases} \quad (29)$$

$$D(\rho) = \begin{cases} Q(\rho) & \text{if } \rho \leq \rho_0, \\ Q(\rho_0) & \text{if } \rho > \rho_0, \end{cases} \quad (30)$$

$$Q(\rho) = q_0 \min\left(\frac{\rho}{\rho_0}, \frac{\rho - \rho_{\text{max}}}{\rho_0 - \rho_{\text{max}}}\right), \quad (31)$$

$q_0 = Q(\rho_0)$ is the maximum flow rate, ρ_{max} is the maximum density (jam density), $x = x_{\text{on}}$ is the location of the farthest upstream cell of the on-ramp merging region of length $L_m = N_m \Delta x$



Modeling Approaches to Traffic Breakdown, Fig. 20 Parameters of Daganzo cell-transmission model (24, 25, 26, 27, 28, 29, 30, and 31) used for numerical simulations of traffic breakdown in Fig. 21: (a) Receiving (R) and sending (D) curves. (b) Model of on-ramp bottleneck. Parameters of the model used in simulations presented in Figs. 21 and 22: $v_{\text{free}} = q_0/\rho_0 = 120$ km/h, $\rho_0 = 15.267$ vehicles/km, $\tau = 0.1$ s, $\Delta x = 10$ m, $L = 20$ km, $N_m = 2$, $t_0 = 7$ min, $x_{\text{on}} = 16$ km, $q_0 = 1832$ vehicles/h

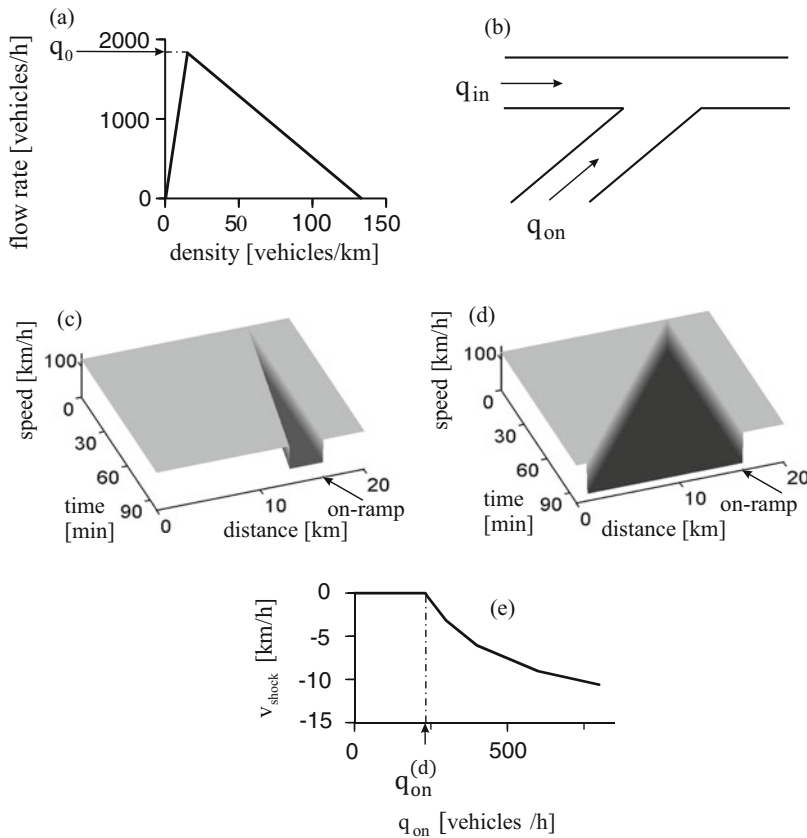
that includes the integer number N_m of cells (Fig. 20b); the flow rate $q_{on}(t)$ is zero at $t < t_0$ and $q_{on}(t) = q_{on}$ is constant at $t \geq t_0$; ρ_{in} is the density in an initial homogeneous free flow on the road, $q_{in} = Q(\rho_{in})$ is the corresponding flow rate.

Achievements of LWR Theory in Description of Traffic Breakdown at Highway Bottleneck

Each of the simulations of congested traffic patterns with the Daganzo model (24, 25, 26, 27, 28, 29, 30, and 31) (Fig. 21) is performed during a limited time interval within which none of the patterns reach the road boundaries, i.e., free flow

conditions remain at the boundaries. The simulations show common qualitative features of traffic breakdown and resulting congested traffic patterns that the LWR theory exhibits at an isolated on-ramp bottleneck (Kerner and Klenov 2006c).

With the use of numerical simulations of Daganzo's cell-transmission model (Daganzo 1994, 1995), one can derive features of traffic breakdown and resulting traffic congestion that the LWR theory exhibits (Fig. 21). Numerical simulations of Daganzo's cell-transmission model presented in Fig. 21 are made on a single-lane road with an on-ramp bottleneck at a given



Modeling Approaches to Traffic Breakdown, Fig. 21 Simulations of traffic breakdown at on-ramp bottleneck on single-lane road in the LWR theory (Kerner 2015b): (a) The fundamental diagram used in simulations of Daganzo's cell-transmission model (24, 25, 26, 27, 28, 29, 30, and 31) (Fig. 20). (b) Qualitative schema of the on-ramp bottleneck. (c, d) Congested traffic at the bottleneck under condition (35) for two

different values of the on-ramp inflow rate q_{on} at a given flow rate q_{in} upstream of the bottleneck: Vehicle speed distributions in space and time that show shock wave emergence and propagation upstream of the on-ramp bottleneck. (e) Dependence of the velocity of the shock wave v_{shock} on the flow rate q_{on} . $q_0 = 1832$ vehicles/h. $q_{in} = 1600$ vehicles/h. $q_{on} = 270$ (c), 600 vehicles/h (d)

flow rate in free flow on the main road upstream of the bottleneck q_{in} and different on-ramp inflow rates q_{on} . We choose the flow rate q_{in} in free flow upstream of the bottleneck that is smaller than q_0 , i.e., we assume that condition

$$q_{in} < q_0 \quad (32)$$

is satisfied.

If the on-ramp inflow rate $q_{on} = 0$, the speed and density are spatially homogeneous. When the on-ramp inflow rate q_{on} increases beginning from zero, the flow rate downstream of the bottleneck $q_{on} + q_{in}$ increases too as long as condition

$$q_{on} + q_{in} < q_0 \quad (33)$$

is satisfied. This is because there is a critical on-ramp inflow rate

$$q_{on}^{(d)} = q_0 - q_{in}. \quad (34)$$

When $q_{on} < q_{on}^{(d)}$, i.e., condition (33) is satisfied, then no congestion occurs at the bottleneck.

However, when $q_{on} > q_{on}^{(d)}$, i.e., condition

$$q_{on} + q_{in} > q_0 \quad (35)$$

is satisfied, which is opposite one to condition (33), then a shock wave of lower speed and larger density appears (Fig. 21c, d). The upstream propagation of the shock wave leads to the widening of congested traffic pattern at the bottleneck (Fig. 21c, d).

The more the flow rate q_{on} exceeds the value $q_{on}^{(d)}$ (34), i.e., the larger the value

$$\Delta q = q_{in} + q_{on} - q_0 > 0, \quad (36)$$

the larger the absolute value of the shock velocity $|v_{shock}|$ (Fig. 21c–e).

Thus, numerical simulations of Daganzo's cell-transmission model show that when traffic breakdown has occurred, i.e., under condition (35), a shock wave propagates upstream of the bottleneck resulting in congested traffic at the bottleneck (Fig. 21c, d). The downstream front of this congested traffic is fixed at the bottleneck

(Fig. 21c, d). Therefore, in accordance with the definition of the synchronized flow phase [S] made in the three-phase theory, this congested traffic belongs to the synchronized flow phase.

Failure of LWR Theory in Explanation of Empirical Nucleation Nature of Traffic Breakdown

Because of above-mentioned achievements of the LWR theory and Daganzo's model in the description of traffic breakdown at a highway bottleneck, a question arises:

- Why does the author make a statement about the *failure* of the LWR theory in the explanation of the empirical nucleation features of traffic breakdown at a highway bottleneck that is the empirical fundamental of transportation science?

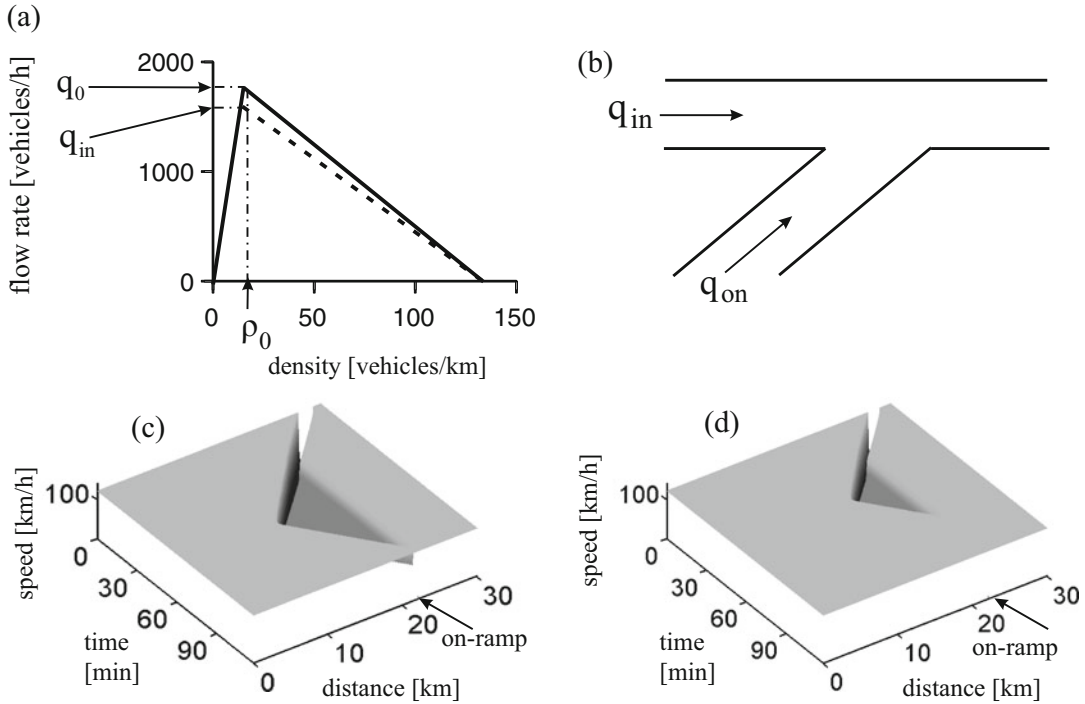
The main disadvantages of the LWR theory in the explanation of traffic breakdown at a bottleneck are as follows:

- (i) The LWR theory cannot show the empirical metastability of free flow with respect to traffic breakdown at the bottleneck.
- (ii) The LWR theory cannot show the empirical induced traffic breakdown at the bottleneck.

In other words, the LWR theory cannot show and explain the evidence of empirical nucleation nature of traffic breakdown at highway bottlenecks (section “[Empirical Proof of Nucleation Nature of Traffic Breakdown \(F→S Transition\)](#).”)

This conclusion is explained in Fig. 22 in which at the same given flow rate q_{in} as that given in Fig. 21 we have chosen smaller q_{on} satisfying condition (33). As follows from Figs. 22c, d, no induced traffic breakdown occurs in free flow at the on-ramp bottleneck.

In simulations, a wide moving jam has been induced at $t = 0$ downstream of the on-ramp bottleneck. As in empirical data (see Fig. 4), while propagating upstream of the on-ramp the jam causes congested traffic at the on-ramp (Fig. 22c, d). However, in contrast with empirical data, in which congested traffic persists at the



Modeling Approaches to Traffic Breakdown, Fig. 22 Continuation of simulations of traffic breakdown at on-ramp bottleneck on single-lane road in the LWR theory (Kerner 2015b): (a) The fundamental diagram (solid curves) is the same as that in Fig. 21a. (b) Qualitative schema of the on-ramp bottleneck taken from Fig. 21b. (c, d) Congested traffic at the bottleneck caused by jam propagation through the bottleneck under condition (33):

Vehicle speed in space and time; the jam width (in the longitudinal direction) at $t = 0$ is equal to $W_J = 2$ (c) and 1.5 km (d). A dashed line in (a) is related to the upstream front of wide moving jams in (c, d), while the jams propagate upstream of the bottleneck. $q_0 = 1832$ vehicles/h, $q_{on} = 120$ vehicles/h, $q_{in} = 1600$ vehicles/h, $\rho_{in} = 13.33$ vehicles/km

bottleneck independent of whether the jam propagates further or the jam dissolves, in the LWR-theory congested traffic always dissolves after the jam dissolution. In the LWR model, congested traffic dissolves independent of the initial jam width W_J (in the longitudinal direction) (Figs. 22c, d) and on the value $q_{on} + q_{in}$ as long as condition (33) is satisfied.

To explain the dissolution of congested traffic, firstly we consider the reason for the dissolution of moving jams that can be seen in Figs. 22c, d. Note that the jam dissolution results from condition $q_{in} < q_0$ (Fig. 22a). Indeed, the downstream jam front moving with the velocity v_g is associated with the line for congested traffic on the fundamental diagram (solid line with negative slope in Fig. 22a). In contrast, when the wide moving jams are upstream of the bottleneck, the line in the

flow–density plane related to the upstream jam front is a dashed line with negative slope shown in Fig. 22a. Due to condition $q_{in} < q_0$, this line (dashed line) is below the line for congested traffic (solid line). Therefore, condition $|v_g| > |v_{up}|$ is satisfied. As a result, the jam width (in the longitudinal direction) decreases over time, i.e., the wide moving jam dissolves upstream of the bottleneck as shown in Figs. 22c, d.

Before the wide moving jam reaches the bottleneck, i.e., when the wide moving jam is still downstream of the bottleneck, the flow rate in the jam outflow $q^{(j)}$ is equal to q_0 . When the jam is upstream of the bottleneck, then within the road region between the bottleneck and the downstream jam front, congested traffic occurs. The flow rate within this congestion is equal to the flow rate in the jam outflow, which decreases to

$q^{(J)} = q_0 - q_{\text{on}}$. This is because in the LWR-model the flow rate downstream of the bottleneck cannot exceed q_0 (Fig. 22a). However, after the wide moving jam has dissolved, the flow rate upstream of the bottleneck decreases to q_{in} resulting in the dissolution of congestion due to condition (33).

- The LWR theory cannot show and explain the empirical evidence of the nucleation nature of empirical traffic breakdown at highway bottlenecks that is the empirical fundamental of transportation science.

Thus, the LWR model cannot show the nucleation nature of traffic breakdown at highway bottlenecks. For this reason, the LWR theory as well as further theoretical approaches based on the LWR theory, like Daganzo's cell transmission model (Daganzo 1994, 1995), so-called N-curves, a macroscopic fundamental diagram (MFD), as well as ITS-applications based on the LWR theory (see references in Sects. 4.3 and 4.11 of the book (Kerner 2017a)) are inconsistent with the nucleation nature of real traffic breakdown at road bottlenecks. Applications of these approaches for an analysis of the effect of ITS on traffic flow, which are widely used by many researchers, can lead to invalid (and sometimes incorrect) conclusions about the ITS performance in real traffic.

- Because the LWR theory is inconsistent with the empirical fundamental of transportation science, theoretical approaches based on the LWR theory cannot be used for a reliable analysis of ITS-applications.

Classical General Motors (GM) Model Class: Free Flow Instability Due to Driver Reaction Time

As long ago as 1958–1961, Herman, Montroll, Potts, Gazis, Chandler, Rothery (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959) and Komentani and Sasaki (Kometani and Sasaki 1958, 1959) assumed that traffic breakdown is associated with an instability of free flow. This instability is related to a finite reaction time of drivers. This driver reaction time is responsible

for the over-deceleration effect discussed in section “Introduction”: If the preceding vehicle begins to decelerate unexpectedly, then owing to the finite driver reaction time the following vehicle starts deceleration with a delay. As a result, if the time delay is long enough, the driver decelerates stronger than it is needed to avoid collisions. In this case, the driver speed can become lower than the speed of the preceding vehicle. If the over-deceleration effect is realized for next following drivers, the wave of vehicle speed reduction appears and increases in amplitude over time in traffic flow leading to a stop of some of the vehicles upstream. This instability should occur when the vehicle density on the fundamental diagram exceeds some critical density denoted $\rho_{\text{cr}}^{(J)}$.

In the car-following model of Herman, Montroll, Potts, Gazis, Chandler, Rothery (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959), which is often called the General Motors (GM) model, the driver reaction time denoted by τ_0 is explicitly used in the vehicle deceleration (acceleration) denoted by $a(t + \tau_0)$, i.e., the vehicle reacts with the time delay τ_0 on any changes in the space gap to the preceding vehicle $g(t)$ and the speed difference $\Delta v(t) = v_{\ell}(t) - v(t)$ (relative speed) between the speed of the preceding vehicle $v_{\ell}(t)$ and the vehicle speed $v(t)$. The GM model reads as follows (Gazis et al. 1961):

$$a(t + \tau_0) = \frac{\Delta v(t)[v(t + \tau_0)]^{m_1}}{T_0[g(t) + d]^{m_2}}, \quad (37)$$

$$a(t) = \frac{dv(t)}{dt}, \quad (38)$$

where d is the vehicle length; T_0 , m_1 , m_2 are constants. By integrating Eqs. 37 and 38, one gets model solutions for steady states related to the fundamental diagram (Gazis et al. 1961)

$$V_0(\rho) = v_0 \left[1 - (\rho/\rho_{\text{max}})^{(m_2-1)} \right]^{(1-m_1)^{-1}}, \quad (39)$$

where v_0 is constant, $\rho_{\text{max}} = 1/d$.

The idea of the GM-model (37) with a driver reaction time is incorporated either explicitly or implicitly in many other traffic flow models

reviewed in (Chowdhury et al. 2000; Cremer 1979; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Whitham 1974; Wiedemann 1974; Wolf 1999).

Newell's Optimal Velocity (OV) Model

At $m_1 = 0$ the GM-model is associated with the Newell optimal velocity (OV) model (Newell 1961)

$$v(t + \tau_0) = V(g(t)). \quad (40)$$

In this case, the OV-relation $v = V(g)$ and the associated fundamental diagram $Q(\rho) = \rho V_0(\rho)$ are determined from Eq. 40 at time-independent space gap g and speed v :

$$v = V(g). \quad (41)$$

Bando et al. Optimal Velocity (OV) Model

In the Bando et al. OV model (Bando et al. 1994, 1995a, b)

$$\frac{dv(t)}{dt} = \frac{V(g(t)) - v(t)}{\tau_0} \quad (42)$$

it is assumed that

$$V(g) = 0.5v_0[\tanh(\beta_0(g - g_0)) + \gamma_0] \quad (43)$$

$\beta_0, \gamma_0, v_0, \tau_0, g_0$ are model parameters.

Treiber et al. Intelligent Driver Model (IDM)

There are many other traffic flow models associated with a generalization of OV-models, which can be written in the form (e.g., Helbing 2001)

$$\frac{dv(t)}{dt} = \phi(v(t), \Delta v(t), g(t)), \quad (44)$$

where the OV-relation $v = V(g)$ and the associated fundamental diagram $Q(\rho) = \rho V_0(\rho)$ are determined from (44) at time-independent space gap g and speeds v, v_ℓ as well as at $v = v_\ell$:

$$\phi(V, 0, g) = 0. \quad (45)$$

In the models (44), a driver reaction time is used only implicitly in a traffic flow model. To these model belongs, for example, the Treiber et al. Intelligent Driver Model (IDM) (Treiber et al. 2000) with

$$\phi = c_1 \left[1 - (v/v_0)^{\delta_0} - (g^*(v, v_\ell)/g)^2 \right], \quad (46)$$

$$g^*(v, v_\ell) = g_1 + c_2(v/v_0)^{1/2} + \tau_1 v + v(v - v_\ell)/2(a_0 b_0)^{1/2}, \quad (47)$$

where $c_1, c_2, \delta_0, g_1, v_0, \tau_1, a_0$, and b_0 are model parameters.

These models exhibit qualitatively the same features of free flow instability and resulting wide moving jam formation as those firstly found in (Kerner and Konhäuser 1994) from an analysis of a version of Payne's macroscopic traffic flow model.

Payne's Macroscopic Traffic Flow Model and its Variants

The Payne-model reads as follows (Payne 1971, 1979):

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial x} = 0, \quad (48)$$

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = \frac{V_0(\rho) - v}{\tau^{(0)}} - \frac{c_0^2}{\rho} \frac{\partial \rho}{\partial x}, \quad (49)$$

where the speed-density relationship $v = V_0(\rho)$ determines the fundamental diagram $Q(\rho) = \rho V_0(\rho)$; $v = v(x, t), \rho = \rho(x, t); \tau^{(0)}, c_0$ are model parameters. To avoid solutions with speed discontinuities, Kühne introduced (Kühne 1991) the viscosity term $\mu \partial^2 v / \partial x^2$ in the Eq. 49 of the Payne-model. In (Kerner and Konhäuser 1994), the Eq. 49 of the Payne-model has been rewritten as the Navier-Stokes equation:

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = \frac{V_0(\rho) - v}{\tau^{(0)}} - \frac{c_0^2}{\rho} \frac{\partial \rho}{\partial x} + \frac{\mu}{\rho} \frac{\partial^2 v}{\partial x^2} \quad (50)$$

with the speed–density relationship

$$V_0(\rho) = c_3 \left[1 + [\exp[(\rho/\rho_{\max}) - c_4]/c_5]^{-1} - c_6 \right], \quad (51)$$

c_i , $i = 3, 4, 5, 6$ are model parameters. There are a variety of other Navier-Stokes-like and gas-kinetic nonlocal traffic flow models, which, as the Payne-model, consist of the vehicle balance and velocity equations (see (Cremer 1979; Papageorgiou 1983) and Section D of (Helbing 2001)). We may call all these macroscopic traffic flow models Payne-like models. This is because all of these macroscopic models show qualitatively the same features of free flow instability and resulting wide moving jam formation in free flow as those firstly found in (Kerner and Konhäuser 1994; Kerner et al. 1995) for a version of the Payne-model (48), (50), (51).

Aw-Rascle Macroscopic Traffic Flow Model

In the Aw-Rascle macroscopic model (Aw and Rascle 2000), instead of the Payne-equation (49), the following velocity equation has been introduced:

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = \frac{V_0(\rho) - v}{\tau^{(0)}} + \rho \frac{dP}{d\rho} \frac{\partial v}{\partial x}, \quad (52)$$

where $P(\rho)$ is a density function. When the same fundamental diagram is chosen in this model as that in the Payne model, then the Aw-Rascle macroscopic model (48), (52) (Aw and Rascle 2000) exhibits also qualitatively the same features of free flow instability and resulting congested pattern formation (Tanga et al. 2007) as those in Payne-like models and other GM model class found earlier in (Helbing et al. 1999; Kerner and Konhäuser 1994; Kerner et al. 1995; Lee et al. 1999).

Wiedemann's Psychophysical Traffic Flow Model

In Wiedemann's psychophysical model (Wiedemann 1974), a driver changes acceleration (deceleration) with a reaction time τ_0 , if some

thresholds (action points) are crossed where the driver changes his or her behavior. These thresholds are usually presented in the relative-speed–space-gap plane for a pair of preceding and follower vehicles. When the relative speed Δv is a large enough negative value and the associated threshold for vehicle deceleration is crossed, the driver decelerates to have the relative speed $\Delta v = 0$ at a desired space gap that depends on the vehicle speed. In the Wiedemann-model, it is assumed that at small values of Δv a driver is not able to recognize whether he or she is slower and faster than the preceding vehicles. For this reason, when the speed of the preceding vehicle is a time-independent one, after reaching the desired space gap, the driver decelerates with a small deceleration before another threshold at a larger space gap is crossed. At this threshold, the driver should recognize that he or she is slower than the preceding vehicle and so accelerates with a small acceleration to the desired space gap, then the driver decelerates with a small deceleration, and so on. Rather than this vehicle motion with small deceleration (acceleration), the dependence of the mentioned desired space gap on speed together with a value of driver reaction time τ_0 determine conditions for traffic flow instability that is qualitatively the same as that in other traffic flow models of the GM model class.

Nagel-Schreckenberg Cellular Automaton (CA) Traffic Flow Model

A different mathematical description for driver reaction time within the framework of the GM model class has been introduced by Nagel and Schreckenberg (1992). In the Nagel-Schreckenberg (NaSch) stochastic cellular automaton (CA) model, in which the time and space are discrete values, the driver reaction time is described through the use of model driver fluctuations. The discrete time is $t = n\tau$, $n = 0, 1, 2, \dots$; τ is the time step and the road is divided into cells of a finite length (see a history of CA traffic flow model development in review (Maerivoet and De Moor 2005)). In the initial version of the NaSch CA model, the cell length has been chosen to be

equal the vehicle length d (Nagel and Schreckenberg 1992). The update rules for vehicle motion in the NaSch CA model can be written as follows

$$v_{n+1} = \max(0, \min(v_{\text{free}}, v_n + 1, g_n) - \xi_n), \quad (53)$$

$$x_{n+1} = x_n + v_{n+1}, \quad (54)$$

$$\xi_n = \begin{cases} 1 & \text{for } r \leq p, \\ 0 & \text{otherwise,} \end{cases} \quad (55)$$

$r = \text{rand}(0, 1)$ denotes a random number uniformly distributed between 0 and 1; $p < 1$; time and space are in the units of τ and d , respectively; g_n is a space gap between two following each other vehicles.

To understand the NaSch-model, firstly note that the space gap g_n determines in Eq. 53 a safe speed: When $v_n > g_n$, then a driver decelerates at time step $n + 1$ because from Eq. 53 it follows that

$$v_{n+1} \leq g_n. \quad (56)$$

This prevents vehicle collisions. Note that the safe speed determined through the space gap g was first introduced by Pipes (1953).

Secondly, the condition

$$v_{n+1} = v_n + 1 \quad (57)$$

means that the vehicle accelerates at time step $n + 1$. However, in the NaSch-model due to model fluctuations with probability p , this acceleration does not occur and instead of Eq. 57 from Eq. 53 it follows that the vehicle maintains its speed:

$$v_{n+1} = v_n. \quad (58)$$

Thus, in this case, model fluctuations simulate a time delay in vehicle acceleration; this time delay is equal to

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p}. \quad (59)$$

Thirdly, let us assume that the vehicle should decelerate at time step $n + 1$ that occurs if $v_n > g_n$.

Then without taking fluctuations into account from Eq. 53 the condition $v_{n+1} = g_n$ would be satisfied. However, due to model fluctuations with probability p , we find $v_{n+1} = g_n - 1$, i.e., the vehicle decelerates stronger than is needed for safety conditions. This is the main idea of the GM-model for the over-deceleration effect discussed above, which should explain traffic flow instability.

In car-following models like the GM-model and different kinds of OV-models, *each* of the vehicles exhibits the same microscopic time delays, which are determined by deterministic rules of model motion (for the models of identical vehicles that are considered here only). In contrast, in the NaSch-model, these driver time delays are simulated through the use of random model fluctuations. As a result, in the NaSch-model, these driver time delays are described as “collective effects” that occur *on average* in traffic flow. The NaSch-model is a theoretical basis for many other CA traffic flow models (see, e.g., (Chowdhury et al. 2000; Knospe et al. 2000, 2002, 2004; Schadschneider et al. 2011) and references there).

Krauß’s Microscopic Stochastic Traffic Flow Model
One of the advantages of the NaSch CA model (53) is the very fast computer simulation times of traffic flow in large networks. A disadvantage of this model in comparison with deterministic traffic flow models is the very large (nonrealistic) vehicle deceleration associated with the model assumption that the safe vehicle speed is given by the space gap g_n (53). We know now that these large model fluctuations are necessary to simulate driver delay times in vehicle acceleration and deceleration. The problem of large speed fluctuations has been solved in the Krauß’s stochastic microscopic model (Krauß et al. 1997), which uses the same ideas for mathematical description of driver delay times in vehicle acceleration and deceleration as those in the NaSch CA model. However, rather than a discrete space of CA models, in the Krauß’s model the continuum space coordinate is used. The Krauß’s model can be written as follows

$$v_{n+1} = \max\left(0, \min\left(v_{\text{free}}, v_n + a\tau, v_n^{(\text{safe})}\right) - \xi_n\right), \quad (60)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (61)$$

$\xi_n = a\tau r$, a is the maximum acceleration, $r = \text{rand}(0, 1)$ denotes a random number uniformly distributed between 0 and 1; the safe speed $v_n^{(\text{safe})} = v^{(\text{safe})}(g_n, v_{\ell, n})$ in (60) is a solution of the Gipps-equation (Gipps 1981)

$$v_n^{(\text{safe})}\tau + X_d\left(v_n^{(\text{safe})}\right) = g_n + X_d(v_{\ell, n}) \quad (62)$$

where $X_d(u)$ is the distance traveled by the vehicle with an initial speed u at a time-independent deceleration b until it comes to a stop; due to discrete model time $t = n\tau$, $n = 0, 1, 2, \dots$ with time step $\tau = 1$ s used in the model (60)

$$X_d(u) = b\tau^2 \left(\alpha\beta + \frac{\alpha(\alpha-1)}{2} \right), \quad (63)$$

where β and α are the fractional and integer parts of $u/b\tau$, respectively.

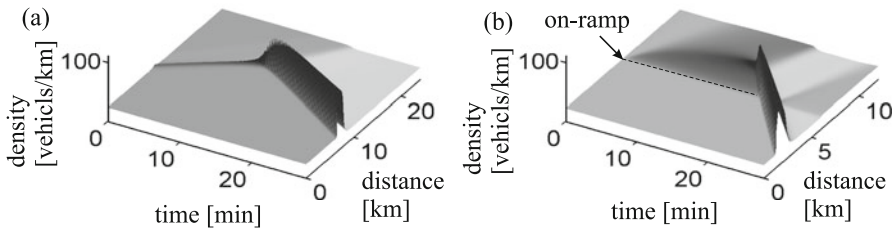
Traffic Breakdown Resulting from Free Flow Instability in GM Model Class: Wide Moving Jam Emergence

What types of congested traffic patterns should occur due to the classical free flow instability in the GM model class? This question has been answered by Kerner and Konhäuser in 1993–1994 from their numerical study of a version of the Payne macroscopic model (48), (50), (51)

(Kerner and Konhäuser 1993; Kerner and Konhäuser 1994). As a result of the instability of free flow at a density that is larger than the critical one $\rho_{cr}^{(J)}$, wide moving jams emerge *spontaneously* in free flow ($F \rightarrow J$ transition for short).

The $F \rightarrow J$ transition exhibits the nucleation nature (Kerner and Konhäuser 1994). This means that at smaller densities in free flow $\rho < \rho_{cr}^{(J)}$, there is a density range (flow rate range) of free flow in which free flow is in a metastable state of free flow with respect to the $F \rightarrow J$ transition. In metastable free flow, a wide moving jam can be excited by a speed (or density) disturbance in an initial homogeneous free flow (Fig. 23a). If an initial amplitude of the growing disturbance is small enough, the $F \rightarrow J$ transition is associated with a “boomerang” behavior of the disturbance (Fig. 23a) (Kerner and Konhäuser 1994): firstly, the disturbance propagates downstream in free flow; then the disturbance comes to a stop strongly growing in its amplitude; as a result, the disturbance begins to propagate upstream; finally, a wide moving jam is forming that propagates upstream, i.e., the jam propagates through the location at which the initial disturbance has occurred.

The same boomerang effect associated with the $F \rightarrow J$ transition has been found, when small amplitude disturbances occur at an on-ramp bottleneck (Fig. 23b) (Kerner et al. 1995). Indeed, due to a speed disturbance that permanently exists at the bottleneck, a wide moving jam emerges spontaneously at the bottleneck in a metastable free flow (Helbing 2001; Helbing et al. 1999; Kerner et al. 1995; Lee et al. 1999; Treiber et al. 2000) (Figs. 23b and 24). At the same flow rate on



Modeling Approaches to Traffic Breakdown, Fig. 23 “Boomerang” behavior of the growing disturbance in metastable free flow by wide moving jam formation. Vehicle density development in space and time.

Simulations of the Payne-like model of Ref. (Kerner and Konhäuser 1994): (a) Homogeneous road (Kerner and Konhäuser 1994). (b) On-ramp bottleneck (Kerner et al. 1995)

the main road upstream of the bottleneck and a larger flow rate to the on-ramp, a sequence of wide moving jams emerges spontaneously (Fig. 24c).

The conclusion of Ref. (Kerner and Konhäuser 1994) that the classical instability of free flow, which should explain traffic breakdown, leads to $F \rightarrow J$ transition is the *general* one for all models of the GM-class, which shows this instability (e.g., Bando et al. 1995a; Helbing 2001; Helbing et al. 1999; Herrmann and Kerner 1998; Lee et al. 1999; Tomer et al. 2000; Treiber et al. 2000; Treiber and Kesting 2013). In particular, $F \rightarrow J$ transitions at the bottleneck for the NaSch CA model (Fig. 24e, f) and for the Wiedemann-model (Fig. 24g, h) show qualitatively the same features of spontaneous wide moving jam emergence as those in a version of the macroscopic Payne-like model of Ref. (Kerner and Konhäuser 1994; Kerner et al. 1995; Lee et al. 1999) (Fig. 24b, c).

In the macroscopic model (48), (50), (51), the safe space gap between vehicles has been simulated through a choice of the fundamental diagram (51) with a very small value $V_0(\rho)$ in the neighborhood of the jam density that exhibits a small absolute value of the derivative $|dV_0/d\rho|$ (Fig. 24a) (Kerner and Konhäuser 1994). Then, at the jam density a vehicle could not almost accelerate from the initial speed $v = 0$ before the space gap to the preceding vehicle increases considerably. Thus, only after a relatively long time delay in vehicle acceleration $\tau_{\text{del}}^{(a)}$ associated with the safe space gap, which considerably exceeds the initial space gap within the jam, could the vehicle escape from the jam. This choice of the fundamental diagram shape simulates also implicitly a much longer time delay in vehicle acceleration $\tau_{\text{del}}^{(a)}$ at speeds that are close to zero $v \approx 0$ within the jam in comparison with considerably shorter time delay in vehicle acceleration at higher speeds. Different time delays in vehicle acceleration $\tau_{\text{del}}^{(a)}$ at $v \approx 0$ and at higher speeds is known as a slow-to-start rule in traffic flow modeling.

By introducing of the slow-to-start rule in the NaSch-model (Barlović et al. 1998), the initial NaSch-model has been further developed to show qualitatively the same features of wide moving jam propagation as those in the Kerner-Konhäuser

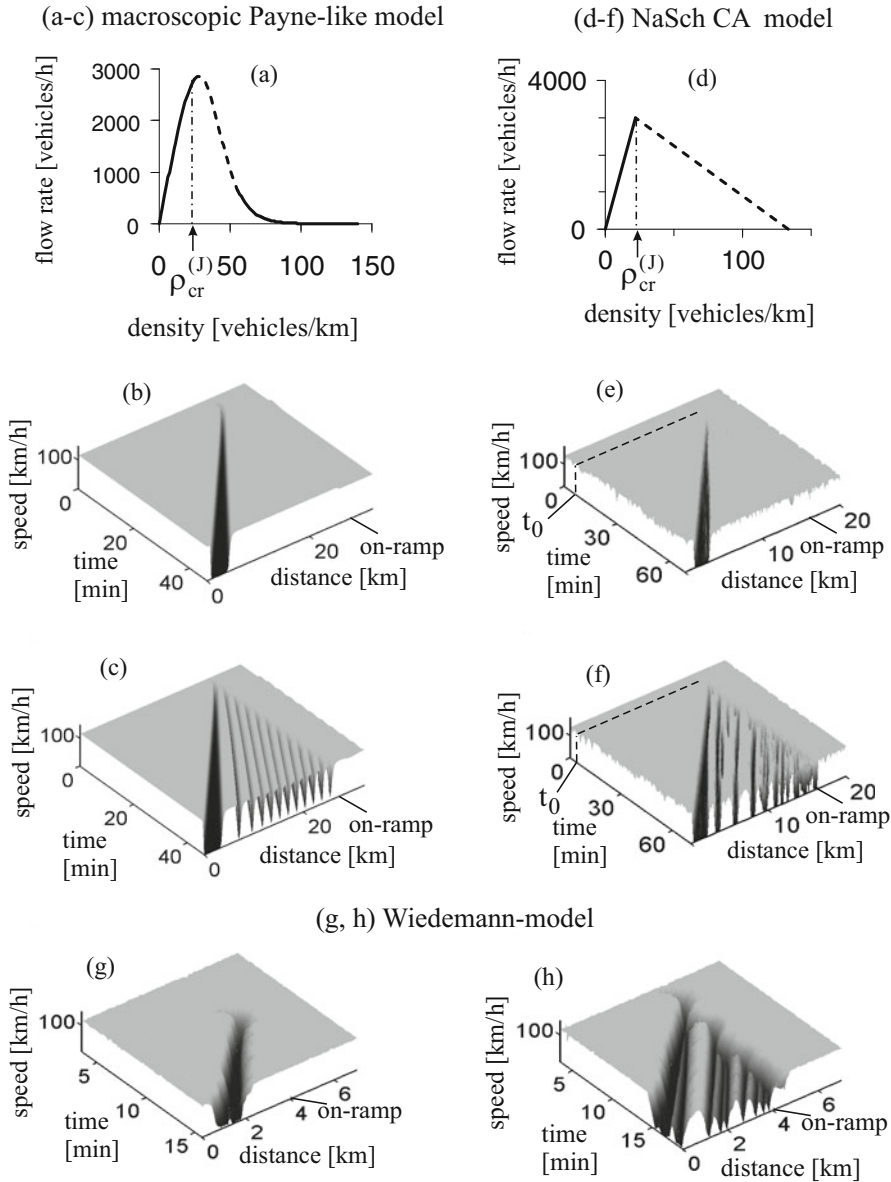
theory of wide moving jams and in empirical data (Kerner 2004a). It must be noted that rather than the implicit simulation of the slow-to-start rule through the use of a special form of the fundamental diagram discussed above, another mathematical idea of simulation of slow-to-start rule firstly introduced by Takayasu and Takayasu (1993) has been used in the NaSch-model (Barlović et al. 1998). Recall that in the NaSch-model the probability of fluctuations p simulates the time delay in acceleration in accordance with (59). Thus, to simulate the slow-to-start rule, this probability has been taken considerably larger at $v = 0$ than at $v > 0$ (Barlović et al. 1998):

$$p(v_n) = \begin{cases} p_1 & \text{for } v_n = 0, \\ p_2 & \text{for } v_n > 0, \end{cases} \quad (64)$$

where p_1, p_2 are constants, $p_1 > p_2$.

Thus, traffic breakdown is explained by an $F \rightarrow J$ transition in all models with the classical free flow instability discussed in this subsection as well as in a huge number of other traffic flow models of the GM-class models (Chowdhury et al. 2000; Helbing 2001; Nagatani 2002; Nagel et al. 2003; Saifuzzaman and Zheng 2014; Treiber and Kesting 2013).

However, this general theoretical result (Helbing 2001; Helbing et al. 1999; Kerner and Konhäuser 1993, 1994; Kerner et al. 1995, 1996; Lee et al. 1999; Mahnke et al. 2005; Nagatani 2002; Nagel et al. 2003; Treiber et al. 2000; Treiber and Kesting 2013) is in contradiction with empirical results; rather than the $F \rightarrow J$ transition, in real traffic flow traffic breakdown is an $F \rightarrow S$ transition (section “[Empirical Proof of Nucleation Nature of Traffic Breakdown \(F→S Transition\)](#)”). In other words, the traffic flow models with the classical free flow instability of the GM-class (Chowdhury et al. 2000; Cremer 1979; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Saifuzzaman and Zheng 2014; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999) cannot be used for a correct description of traffic breakdown at highway bottlenecks.



Modeling Approaches to Traffic Breakdown, Fig. 24 Traffic breakdown at on-ramp bottleneck in the GM-model class with classical free flow instability (Helbing 2001; Kerner 2009c; Kerner et al. 1995; Lee et al. 1999; Treiber et al. 2000; Treiber and Kesting 2013): Wide moving jam formation in free flow in a Payne-like model of Ref. (Kerner and Konhäuser 1994) (a–c), in the NaSch CA model (d–f), and in the Wiedemann-model (g, h). (a, d) – The fundamental diagrams of the Payne-like model (with the same model parameters as those used in Fig. 23a) and of the NaSch

CA model (d) (Kerner and Konhäuser 1994); dashed parts of the diagrams are related to unstable states (b, c, e, f–h) – Average speed in space and time. In (g, h), simulations have been made with tool VISSIM 4.10–10 based on the Wiedemann-model; simulation parameters: flow rate in free flow on the main single-lane road upstream of the bottleneck $q_{in} = 2750$, flow rate to the on-ramp $q_{on} = 200$ vehicles/h for (g) and $q_{in} = 2500$, $q_{on} = 400$ vehicles/h for (h), length of on-ramp merging region is 300 m, maximum speed in free flow is within the range 119–120 km/h, passenger vehicles only

Summary of Achievements of Classical Traffic Flow Models

We can summarize the main achievements of the classical traffic flow models as follows.

1. The LWR-theory (Daganzo 1994, 1995; Lighthill and Whitham 1955; Richards 1956) can show a transition to traffic congestion whose downstream front is fixed at a bottleneck.
2. Traffic flow models in the framework of the GM model class can explain at least the following empirical features of traffic congestion:
 - Characteristic parameters of wide moving jams and line J (Kerner and Konhäuser 1994).
 - A broad spread of traffic data in the flow–density plane associated with congested traffic (e.g., Helbing 2001; Treiber and Kesting 2013).
3. Ideas of classical models for description of the driver reaction time and over-deceleration effect, the slow-to-start rule, simulations of driver time delays through model fluctuations, and many others, which were introduced in earlier models and theories by Herman, Montroll, Potts, Rothery, Chandler, Gazis (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959), Komentani and Sasaki (1958), Newell (1961, 2002), Gipps (1981), Payne (1971, 1979), Nagel, Schreckenberg, Schadschneider, and co-workers (Barlović et al. 1998; Nagel and Schreckenberg 1992), Bando, Sugiyama, and colleagues (1995a), Takayasu and Takayasu (1993), Krauß et al. (1997), Nagel et al. (1998), Nagatani (1998, 1999), Nagatani and Nakanishi (1998) and by many other groups (see references in (Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Elefteriadou 2014; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Maerivoet and De Moor 2005; Mahnke et al. 2005, 2009; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Saifuzzaman and Zheng 2014; Schadschneider et al. 2011; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999)), are also very

important for traffic flow models in the framework of the three-phase theory (see the books (Kerner 2004a, 2009c, 2017a)).

Why Are Generally Accepted Classical Two-Phase Traffic Flow Models Inconsistent with Features of Real Traffic?

Because of great achievements of generally accepted classical traffic flow models discussed in sections “Achievements of LWR Theory in Description of Traffic Breakdown at Highway Bottleneck” and “Summary of Achievements of Classical Traffic Flow Models,” a question arises:

- Why does the author state that the generally accepted classical traffic flow theories and models are not applicable for a reliable description of traffic breakdown, highway capacity, traffic control, dynamic traffic assignment, and optimization of real traffic and transportation networks?

The failure of the LWR-theory, the Daganzo’s cell transmission model, traffic flow models of the GM model class as well as other two-phase traffic flow models is explained as follows:

1. The LWR-model and the Daganzo’s cell transmission model failed because these models cannot show the metastability of free flow with respect to traffic breakdown as well as induced traffic breakdown at a highway bottleneck observed in real traffic (section “Empirical Proof of Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition).”)
2. Traffic flow models of the GM model class failed because the models cannot show the metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck as observed in all real field traffic data (section “Empirical Proof of Nucleation Nature of Traffic Breakdown ($F \rightarrow S$ Transition).”). Recall that the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck explains the nucleation nature of traffic breakdown at highway bottlenecks that is the empirical fundamental of transportation science.

- The classical traffic flow models failed because they are inconsistent with the empirical fundamental of transportation science – the nucleation nature of traffic breakdown at highway bottlenecks.

These critical conclusions are independent of the choice of model parameters. For example, due to an appropriate choice of model parameters in a traffic flow model of the GM model class, the classical traffic flow instability of the GM model class can be realized in the model only at a vehicle density that is a larger one than the density associated with the maximum on the fundamental diagram of the model. In this case, the traffic flow model exhibits features of both the LWR model and a model of the GM model class. However, such a “combined” traffic flow model cannot also show the metastability of free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at the bottleneck (for more detailed explanations, see section 4.8 of the book (Kerner 2017a)).

As mentioned in section “[Failure of Classical Modeling Approaches to Traffic Breakdown at Highway Bottlenecks](#),” none of classical traffic-flow theories and models (see, e.g., reviews (Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Eleftheriadou 2014; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Maerivoet and De Moor 2005; Mahnke et al. 2005, 2009; May 1990; Nagatani 2002; Nagel et al. 2003; Papageorgiou 1983; Saifuzzaman and Zheng 2014; Schadschneider et al. 2011; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999)) incorporates the metastability of free flow with respect to an $F \rightarrow S$ transition at a bottleneck.

- To distinguish three-phase traffic flow models in the framework of the three-phase theory discussed below from classical traffic flow theories and models (in particular, the LWR model and traffic flow models of the GM model class as well as any combinations of the classical theories and models), the classical traffic flow theories and models can be considered “two-phase traffic flow models.”

- Thus, in the article the term *a two-phase traffic flow model* is a synonym of the terms *a classical traffic flow model* and *a classical traffic flow theory*.

Main Prediction of Three-Phase Theory

Explanation of Empirical Nucleation Nature of Traffic Breakdown

The empirical nucleation nature of traffic breakdown (metastability of free flow with respect to an $F \rightarrow S$ transition) at highway bottlenecks discussed in section “[Empirical Proof of Nucleation Nature of Traffic Breakdown \(\$F \rightarrow S\$ Transition\)](#)” was understood in 1996–2002 based on an analysis of a number of real field traffic data measured over many days (and years) on several German highways (see (Kerner 1998a, b, 1999a, 2000, 2001, 2002a, b; Kerner and Rehborn 1996a, b, 1997) and references in (Kerner 2004a)). The understanding of these empirical data has led to the formulation by the author of the three-phase theory (Kerner 1998a, b, 1999a, b, c).

- The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown at highway bottlenecks.

The three-phase theory consists of several hypotheses resulting from a spatiotemporal analysis of real field traffic data measured during many years of traffic observations on different highways in a variety of countries. The hypotheses of the three-phase theory firstly formulated in (Kerner 1998a, b, c, 1999a, b, c) have been considered in details in the books (Kerner 2004a, 2009c). A summary of the hypotheses of the three-phase theory can also be found in section 1.9 of the book (Kerner 2017a).

Here, we consider briefly only hypotheses of the three-phase theory that explain the empirical nucleation nature of traffic breakdown. The metastability of free flow with respect to an $F \rightarrow S$ transition has been explained in the three-phase theory through the following two hypotheses (Kerner 1998a, b, 1999a, 2004a, 2009c):

- The metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck is caused by a competition between two opposing tendencies – the tendency towards free flow due to the *over-acceleration effect* (driver over-acceleration) and the tendency towards synchronized flow due to the *speed adaptation effect* (driver speed adaptation).
- The tendency towards free flow due to the over-acceleration effect is associated with a traffic flow instability in synchronized flow caused by a discontinuous character of a density (flow-rate) function of the probability of driver over-acceleration. The discontinuous character of driver over-acceleration is associated with the existence of a driver time delay in over-acceleration.
- This instability in synchronized flow causes a growing speed wave of a local increase in the speed in synchronized flow. The uninterrupted growth of the speed wave leads to a phase transition from synchronized flow (S) to free flow (F) ($S \rightarrow F$ transition). For this reason, we call this instability as an $S \rightarrow F$ instability.

In other words, it is assumed in the three-phase theory (Kerner 2004a) that due to a driver time delay in over-acceleration, the probability of driver over-acceleration in car following exhibits a discontinuous character. The discontinuous character of the over-acceleration leads to the $S \rightarrow F$ instability, i.e., to the emergence of a growing wave of a local increase in the speed in synchronized flow. This traffic flow instability ($S \rightarrow F$ instability) can be considered the result of the over-acceleration effect that determines the tendency towards free flow within a local speed increase in synchronized flow.

In the three-phase theory, it is assumed that a local speed disturbance in free flow at a bottleneck is a nucleus required for traffic breakdown, if the over-acceleration effect within this speed disturbance is on average weaker than the speed adaptation effect. Otherwise, if within a local speed disturbance the over-acceleration effect is on average stronger than the speed adaptation effect, the local speed disturbance decays over time, i.e., this

disturbance is not a nucleus required for traffic breakdown.

The main prediction of the three-phase theory that distinguishes the three-phase theory from any earlier traffic flow theories and models is as follows:

- The three-phase theory predicts that there is an $S \rightarrow F$ instability of synchronized flow, which exhibits the nucleation nature. The nucleation nature of the $S \rightarrow F$ instability governs the nucleation nature of the $F \rightarrow S$ transition (traffic breakdown) at highway bottlenecks.
- The nucleation nature of the $S \rightarrow F$ instability has a sense for *none* of the earlier traffic flow theories and models. For this reason, the three-phase theory is incommensurable with any earlier traffic flow theories and models.

The term “incommensurable” has been introduced by Kuhn (Kuhn 2012) to explain a paradigm shift in a scientific field.

Criticism of Three-Phase Traffic Theory

The main criticism on the three-phase theory made during last years, in particular in the works by Daganzo et al. (1999) and by Helbing, Treiber et al. (Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010) can be formulated as follows:

- The three-phase theory is not needed because generally accepted traffic flow theories also succeed in simulating of the most essential empirical features of traffic flow described by the three-phase theory.

The authors of these works believe (Daganzo et al. 1999; Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010) that the properly tuned two-phase traffic flow models, based on extensive simulations can show the same spatiotemporal traffic features as three-phase traffic flow models.

To prove their statements, Helbing, Treiber et al. (Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010) are arguing that the

well-defined empirical features of the spatiotemporal congested patterns, as well as empirical 2D scattering of the flow-density plot, can be captured by two-phase traffic flow models. In particular, two-phase traffic flow models can indeed show 2D scattering of points in the flow-density plot as well as some spatiotemporal congested patterns that look similar to results of empirical observations.

- However, rather than empirical features of the spatiotemporal congested patterns and 2D empirical scattering of the flow-density plot, the metastability of free flow at a highway bottleneck with respect to an $F \rightarrow S$ transition is needed to use as the basic criterion for a comparison of the classical traffic flow models (Daganzo et al. 1999; Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010) with three-phase traffic flow models (Kerner 2004a).

Indeed, the main prediction of the three-phase theory is the explanation of traffic breakdown through the metastability of free flow at a bottleneck with respect to an $F \rightarrow S$ transition (section “Explanation of Empirical Nucleation Nature of Traffic Breakdown”). Such a comparison has *not* been made either by Daganzo et al. (1999) or by Helbing, Treiber et al. (Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010). The reason for the invalid criticism of the three-phase theory made by Daganzo et al. (1999) and Helbing, Treiber et al. (Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010) is that none of the traffic flow models of Daganzo et al. (1999), and none of the traffic flow models of Helbing, Treiber et al. (Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010) is able to show the empirical fundamental of transportation science – the metastability of free flow at a highway bottleneck with respect to an $F \rightarrow S$ transition.

Thus, in contrast with the three-phase theory, none of the classical traffic theories and two-phase traffic flow models can show the metastability of free flow with respect to an $F \rightarrow S$ transition at a

highway bottleneck. Nevertheless, the following question can arise:

- Why does the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck, which is the reason for the three-phase theory, exhibit the *fundamental importance* for transportation science?

The metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck is needed to explain the empirical nucleation nature of traffic breakdown that is the empirical fundamental of transportation science. For this reason, traffic and transportation theories, which are not consistent with the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway bottlenecks, cannot be applied for the development of reliable traffic control, dynamic traffic assignment, and organization as well as other ITS-applications in traffic and transportation networks.

This explains why *neither* empirical features of spatiotemporal congested patterns, *nor* the empirical 2D scattering of the flow-density plot should be considered as the basic empirical traffic phenomena for a comparison of two-phase and three-phase traffic flow models as made by Daganzo, Helbing, Treiber et al. (Daganzo et al. 1999; Helbing et al. 2009; Schönhof and Helbing 2007, 2009; Treiber et al. 2010). This is because the fundamental difference between two-phase and three-phase traffic flow models is associated with the explanation of the empirical nature of traffic breakdown at a highway bottleneck:

- *None* of the two-phase traffic flow models can simulate and explain the empirical metastability of free flow at a highway bottleneck with respect to an $F \rightarrow S$ transition as observed in real field traffic data.
- The explanation of the empirical metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck is the main reason for the three-phase theory (Kerner 2004a). Therefore, the statement “the three-phase theory is not needed” is invalid.

Competition of Driver Speed Adaptation with Driver Over-Acceleration in Three-Phase Theory

Two-Dimensional (2D) Traffic Flow States

For a qualitative consideration of the speed adaptation effect, we introduce the term a *steady state of synchronized flow*:

- A steady state of synchronized flow is a *hypothetical* state of synchronized flow of identical vehicles and drivers in which all vehicles move with the same time-independent speed v and have the same space gap g , i.e., this synchronized flow is homogeneous in time and space. In any steady state of synchronized flow, a steady vehicle speed $v > 0$.

In the three-phase theory, the following hypothesis about two-dimensional (2D) steady states of synchronized flow has been introduced (Kerner 1998a, b, 1999a):

- Steady states of synchronized flow cover a two-dimensional (2D) region in the flow–density plane (Fig. 25a).

This hypothesis means that alone interactions between identical vehicles and drivers are responsible for 2D-steady states in synchronized flow in accordance with the following driver behavioral assumptions (see Sect. 4.3 and 8.6 of the book (Kerner 2004a) for detailed explanations): Even if the speed difference between vehicles is negligible,

a driver recognizes whether the space gap to the preceding vehicle becomes larger or smaller. This is true for each synchronized flow speed within a finite space-gap range. Therefore, we can assume that a given steady speed can be related to the infinite number of space gaps; respectively, a given space gap between vehicles can be related to the infinite number of steady speeds (Fig. 25b).

The upper boundary of the 2D region of the steady states of synchronized flow (labeled by S_{upper} in Fig. 25a) is determined by a *safe space gap* denoted by g_{safe} (Fig. 25b). The safe space gap is associated with a safe time headway denoted by τ_{safe} that is determined from the equation:

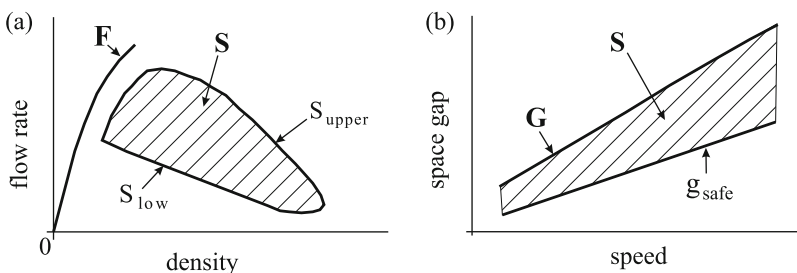
$$g_{\text{safe}}(v) = v\tau_{\text{safe}}(v). \quad (65)$$

The lower boundary of the 2D-region for steady states of synchronized flow in the flow–density plane (labeled by S_{low}) is determined by a *synchronization space gap* between vehicles denoted by G (Fig. 25b). The synchronization space gap is associated with a synchronization time headway denoted by τ_G that is determined from the equation:

$$G(v) = v\tau_G(v). \quad (66)$$

In Eqs. 65 and 66, we have taken into account that in a general case the safe time headway τ_{safe} and/or the synchronization time headway τ_G can be speed functions.

In accordance with this hypothesis of the three-phase theory (Kerner 1998a, b, 1999a, b, 2004a),



Modeling Approaches to Traffic Breakdown, Fig. 25 Hypothesis about steady states of synchronized flow in the three-phase theory (Kerner 1998a, b, 1999a): (a) Qualitative representation of free flow states (F) and 2D

steady states of synchronized flow (dashed region S) on a multi-lane road in the flow–density plane. (b) A part of the 2D steady states of synchronized flow shown in (a) in the space-gap–speed plane (dashed region S)

at a given steady speed v in synchronized flow, there are the *infinite number* of space gaps g within the range

$$g_{\text{safe}}(v) \leq g \leq G(v) \quad (67)$$

at which a driver can move with this steady, i.e., time-independent speed v . Conditions (67) are equivalent to

$$\tau_{\text{safe}}(v) \leq \tau^{(\text{net})} \leq \tau_G(v), \quad (68)$$

where $\tau^{(\text{net})}$ is time headway (net time gap) to the preceding vehicle.

Respectively, at a given space gap g between vehicles in synchronized flow, there are the *infinite number* of vehicle steady speeds v within the range

$$v_G(g) \leq v \leq v_{\text{safe}}(g), \quad (69)$$

where the speed $v_G(g)$ is a solution of the equation

$$g = G(v_G) \quad (70)$$

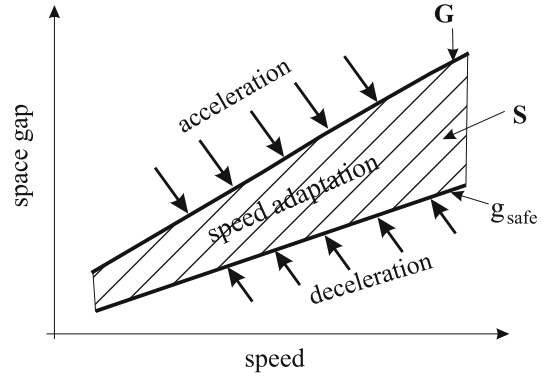
and the safe speed $v_{\text{safe}}(g)$ is a solution of the equation

$$g = g_{\text{safe}}(v_{\text{safe}}). \quad (71)$$

It can be assumed that when the space gap g to the preceding vehicle becomes larger than the synchronization space gap G , the vehicle accelerates (labeled by “acceleration” in Fig. 26). When the space gap g becomes smaller than the safe space gap g_{safe} , the vehicle decelerates (labeled by “deceleration” in Fig. 26).

A more detailed consideration of Figs. 25b and 26 together with conditions (67) and (69) shown in Fig. 27 allows us to understand some of the consequences of the hypothesis of the three-phase theory about 2D-steady states of synchronized flow.

In Fig. 27c, we assume that all vehicles move at a given steady speed (dashed line $v = \text{const}$ in Fig. 27a) and the speed difference



Modeling Approaches to Traffic Breakdown, Fig. 26 Qualitative explanation of car-following in the three-phase theory (Kerner 1998a, b, 1999a, 2004a, 2009c): A vehicle accelerates at $g > G$ (labeled by “acceleration”) and decelerates at $g < g_{\text{safe}}$ (labeled by “deceleration”), whereas under condition (67) the vehicle adapts its speed to the speed of the preceding vehicle without caring what the precise space gap is (labeled by “speed adaptation”). A dashed region of synchronized flow is taken from Fig. 25b

$$\Delta v = v_\ell - v \quad (72)$$

between the speed of the preceding vehicle v_ℓ and the vehicle speed v is zero: $\Delta v = 0$. Now under conditions

$$v = \text{const}, \Delta v = 0 \quad (73)$$

we consider steady states of synchronized flow at different space gaps. As long as conditions (67) and (73) are satisfied, in accordance with the hypothesis of the three-phase theory about 2D-steady states of synchronized flow, the vehicle acceleration (deceleration) is equal to zero (Fig. 27c). However, when the space gap between vehicles is larger than the synchronization gap

$$g > G(v), \quad (74)$$

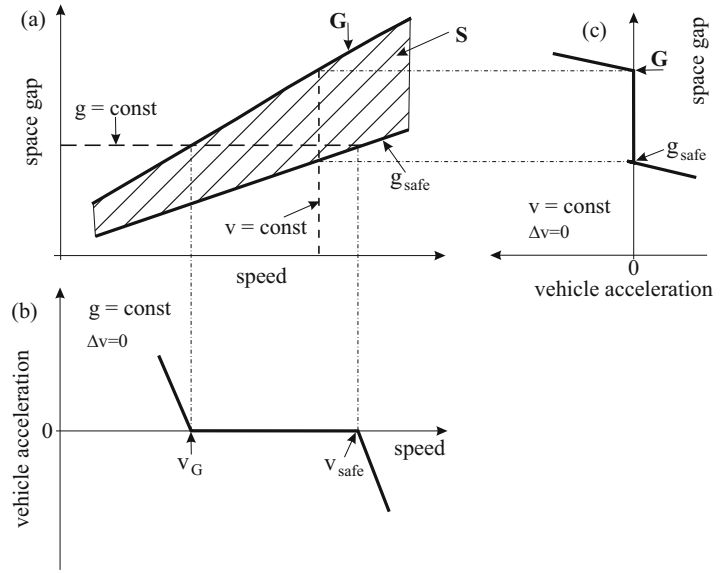
conditions (67) are not valid anymore and the vehicle accelerates. Respectively, when the space gap between vehicles is smaller than the safe gap

$$g < g_{\text{safe}}(v), \quad (75)$$

conditions (67) are also not valid anymore and the vehicle decelerates (Fig. 27c).

Modeling Approaches to Traffic Breakdown,

Fig. 27 Qualitative explanations of formulas (67) and (69): (a) 2D-region of steady states of synchronized flow taken from Fig. 25b with dashed lines $g = \text{const}$ and $v = \text{const}$. (b) Qualitative dependence of vehicle acceleration at $g = \text{const}$ and $\Delta v = 0$ (76). (c) Qualitative dependence of vehicle acceleration at $v = \text{const}$ and $\Delta v = 0$ (73)



In Fig. 27b, we assume that all vehicles move at a given space gap (dashed line $g = \text{const}$ in Fig. 27a). Now under conditions

$$g = \text{const}, \Delta v = 0 \quad (76)$$

we consider steady states of synchronized flow at different vehicle speeds. As long as conditions (69) and (76) are satisfied, in accordance with the hypothesis of the three-phase theory about 2D-steady states of synchronized flow, the vehicle acceleration (deceleration) is equal to zero (Fig. 27b). When the steady speed v is smaller than the steady speed at the synchronization gap

$$v < v_G(g), \quad (77)$$

conditions (69) are not valid anymore and the vehicle accelerates. Contrarily, when the steady speed v is larger than the safe speed

$$v > v_{\text{safe}}(g), \quad (78)$$

conditions (69) are also not valid anymore and the vehicle decelerates (Fig. 27b).

driver decelerates within the synchronization space gap G adapting the speed to the speed of the preceding vehicle (Fig. 26). This speed adaptation within the synchronization gap is associated with the 2D-region of steady states of synchronized flow given by conditions (67) (Fig. 25b). It should be noted that when the vehicle speed v and the speed of the preceding vehicle v_ℓ are time-functions, then rather than conditions (67) and (68) for steady states, we should consider, respectively, conditions

$$g_{\text{safe}} \leq g \leq G \quad (79)$$

and

$$\tau_{\text{safe}} \leq \tau^{(\text{net})} \leq \tau_G \quad (80)$$

in which the synchronization space gap G and time headway τ_G as well as the safe space gap g_{safe} and time headway τ_{safe} can be complex dynamic variables depending, for example, on $g(t)$, $v(t)$, and $v_\ell(t)$.

In general, the speed adaptation effect is defined as follows:

Speed Adaptation Effect Within Dynamic 2D-States of Synchronized Flow

When a driver approaches a slower moving preceding vehicle and the driver cannot pass it, the

- The speed adaptation effect is the adaptation of the vehicle speed to the speed of the preceding vehicle at any space gap within the space gap range between the synchronization space gap

and the safe space gap. At each vehicle speed within synchronized flow, while adapting the vehicle speed to the speed of the preceding vehicle, a driver can make an arbitrary choice in the space gap that satisfies conditions (79): The driver accepts different spaces gaps at different times and does not control a fixed space gap to the preceding vehicle. In other words, within the space gap range (79), the driver speed adaptation occurs without caring what the precise space gap to the preceding vehicle is.

If the preceding vehicle moves at a time-independent synchronized flow speed, and if we neglect fluctuations, the following driver moves with the speed of the preceding vehicle (the vehicle acceleration is equal to zero, Figs. 27b, c) at any space gap from the space gap range (79), i.e., at any time headway from the time headway range (80). This is due to the driver behavioral assumption associated with driver interactions in synchronized flow that at a given speed in steady states of synchronized flow, the driver can make an arbitrary choice in the space gap to the preceding vehicle within the gap range (79). Thus, under condition (79) (or condition (80)), drivers do not try to reach a particular space gap (particular time headway) to the preceding vehicle, but adapt the speed while keeping the space gap (time headway) in a range.

There can be a number of mathematical formulations for the description of the speed adaptation effect of the three-phase theory. One of the simplest ones is related to a simple equation

$$A = K\Delta v \text{ at } g_{\text{safe}} \leq g \leq G, \quad (81)$$

where $A = dv/dt$ is the vehicle acceleration (deceleration),

$$\Delta v = v_\ell - v \quad (82)$$

is the speed difference between the speed of the preceding vehicle v_ℓ and the vehicle speed v , K is a constant coefficient ($K > 0$). Equation 81 that describes the speed adaptation effect within 2D-states of synchronized flow has firstly been applied in the Kerner-Klenov microscopic

stochastic model (Kerner and Klenov 2002, 2003, 2009b) and the KKW CA model (Kerner et al. 2002) as well as in the Kerner-Klenov ATD (acceleration time delay) microscopic deterministic three-phase model (Kerner and Klenov 2006c).

When the driver is slower than the preceding vehicle $v < v_\ell$ then from Eqs. 81 and 82, we get $A > 0$, i.e., in accordance with the speed adaptation effect a driver accelerates adapting the speed to the speed of the preceding vehicle. Contrarily, when the driver is faster than the preceding vehicle $v > v_\ell$ then from Eqs. 81 and 82 we get $A < 0$, i.e., the driver decelerates adapting the speed to the speed of the slower moving preceding vehicle. In accordance with Eq. 81, the speed adaptation occurs at any space gap within the space gap range between the synchronization space gap G and the safe space gap g_{safe} : The driver accepts different spaces gaps at different times and does not control a fixed space gap to the preceding vehicle.

Equation 81 has also been applied in automated driving (called also as self-driving or automatic driving, or else autonomous driving) based on the three-phase theory that has been assumed by the author in inventions (Kerner 2004b, 2007e, f) (see Sect. 9.6 in the book (Kerner 2009c) and chapter “► Mathematical Probabilistic Approaches to Traffic Breakdown” of the book (Kerner 2017a)). In this case, the acceleration (deceleration) of the automated driving vehicle is equal to $a = K_{\Delta v}\Delta v$ at $g_{\text{safe}} \leq g \leq G$, where $K_{\Delta v}$ is a dynamic coefficient of an automated driving vehicle ($K_{\Delta v} > 0$). We can see that this equation coincides with Eq. 81 for manual driving vehicles. The effect of automated driving in the framework of the three-phase theory on mixed traffic flow consisting of automated driving and manual driving vehicles has been studied in (Kerner 2017c).

It must be emphasized (Kerner 1998a, b, 1999a, 2004a) that 2D-steady states of synchronized flow are only some *hypothetical flow states*, which cannot be exactly found and, therefore, measured in real synchronized flow. This is at least because of fluctuations, which are always present in real traffic flow. Moreover, we should mention that moving during a long enough time interval the driver can arbitrarily change the space gap within the range (79): She/he moves usually at a particular space gap only during a finite time

interval changing to the moving at another space gap within the range (79) during another time interval, and so on. If we denote random fluctuations in driver acceleration (deceleration) in 2D-dynamic states of synchronized flow by $\varsigma(t)$, then Eq. 81 can be written as

$$A = K\Delta v + \varsigma \text{ at } g_{\text{safe}} \leq g \leq G. \quad (83)$$

This simple formulation of the speed adaptation effect (83) was firstly introduced in 2002 in the Kerner-Klenov microscopic stochastic three-phase traffic flow model (Kerner and Klenov 2002) (see formula (11.54) of the book (Kerner 2009c)), and it has been generalized for this model in (Kerner and Klenov 2009b) (see section “Speed Adaptation Effect”). It should be noted that in Eq. 83 the safe space gap g_{safe} and the synchronization space gap G can depend on the vehicle speed v and the speed of the preceding vehicle v_ℓ , which can be complex functions of time t .

In other words, even at a given speed in synchronized flow there are always dynamic transitions over time between different 2D-states of synchronized flow (Kerner 1998a, b, 1999a, 2004a). The dynamic 2D-states of synchronized flow are nonhomogeneous in space and time. However, these dynamic synchronized flow states can approximately correspond to the hypothetical 2D-steady states of synchronized flow.

In the three-phase theory (Kerner 1998a, b, 1999a, 2004a), it is assumed that features of instabilities and resulting phase transitions that can occur in the dynamic 2D-states of synchronized flow should be qualitatively the same as those in 2D-steady states of synchronized flow. This assumption of the three-phase theory is also related to an $S \rightarrow F$ instability. This emphasizes the importance of a consideration of 2D-steady states of hypothetical synchronized flow for the understanding of the nature of traffic breakdown as well as other traffic flow phenomena in dynamic nonhomogeneous 2D-states of real synchronized flow.

This explains the hypothesis of the three-phase theory about 2D-steady states of synchronized flow (Kerner 1998a, b, 1999a, 2004a) as follows.

- At each vehicle speed within synchronized flow, while adapting the vehicle speed to the speed of the preceding vehicle, a driver can make an arbitrary choice in the space gap (time headway) that satisfies conditions (79) (condition (80)): The driver accepts different space gaps (different time headway) at different times and does not control a fixed space gap (fixed time headway) to the preceding vehicle. This driver speed adaptation effect within 2D-states of synchronized flow is independent of whether steady states of hypothetical synchronized flow or dynamic nonhomogeneous states of real synchronized flow are considered.

Driver Over-Acceleration

When condition (79) is satisfied, a vehicle accelerates usually when it is slower than the preceding vehicle, and decelerates when it is faster than the preceding vehicle. However, under condition (79) the vehicle can also accelerate even if the vehicle speed $v(t)$ is *not* smaller than the speed of the preceding vehicle $v_\ell(t)$, i.e., when condition

$$v(t) \geq v_\ell(t) \quad (84)$$

is satisfied. Such vehicle acceleration can be considered *vehicle over-acceleration* (Kerner 2004a). In general, the over-acceleration effect is defined as follows:

- The over-acceleration effect is driver maneuver leading to a higher speed from initial car-following at a lower speed occurring under condition (79).

It must be noted that in real free flow the speed adaptation and over-acceleration effects appear usually in their dynamic competition within a local disturbance in free flow at a bottleneck. For this reason, a separate consideration of the over-acceleration effect made here is a simplification of the reality (see section “ $S \rightarrow F$ Instability at Bottleneck” below).

Accordingly to a hypothesis of the three-phase theory, at a given vehicle density the probability of over-acceleration P_{OA} is larger in free flow than

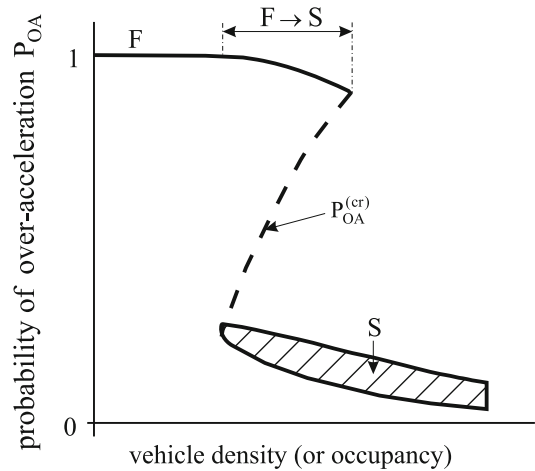
in synchronized flow. This assumption leads to the following hypothesis of the three-phase theory about a discontinuous character of probability of driver over-acceleration (Kerner 1999a, b, 2004a):

- Probability of over-acceleration exhibits a *discontinuous* character. In particular, the probability is a Z-shaped density function (Fig. 28): At a given density, there is a *drop* in probability of over-acceleration P_{OA} , when free flow transforms into synchronized flow.

This hypothesis is explained by a driver behavioral assumption made in the three-phase theory that there should be a driver time delay in over-acceleration. Indeed, while moving in synchronized flow, a driver has to wait while searching for the opportunity to accelerate to a higher speed in free flow.

These features of over-acceleration lead to a conclusion that when a driver moves initially in synchronized flow, an $S \rightarrow F$ instability caused by over-acceleration should occur. This $S \rightarrow F$ instability should cause a growing wave of the speed increase in synchronized flow. The uninterrupted growth of this wave leads to a transition from synchronized flow to free flow ($S \rightarrow F$ transition). In its turn, due to this $S \rightarrow F$ instability, there should be a drop in the probability of over-acceleration (Fig. 28). This explains the above-mentioned hypothesis about a discontinuous character of over-acceleration.

For a qualitative explanation of the $S \rightarrow F$ instability, we assume that a driver moves initially in synchronized flow. If a driver increases the speed due to over-acceleration, the probability of over-acceleration for the following drivers increases. Therefore, the subsequent over-acceleration of the following drivers becomes more probable. Due to the time delay in driver over-acceleration, it takes some time for the development of this over-acceleration effect. In other words, the $S \rightarrow F$ instability caused by over-acceleration should be associated with a growing wave of a local increase in the vehicle speed in synchronized flow. In section “ $S \rightarrow F$ Instability at Bottleneck,” we will show that this assumption



Modeling Approaches to Traffic Breakdown, Fig. 28 Qualitative Z-shaped density function of probability of over-acceleration (Kerner 1999a, b). F – free flow, S – synchronized flow, dashed curve between states F and S is related to probability of over-acceleration $P_{OA}^{(cr)}$ within critical nuclei for traffic breakdown ($F \rightarrow S$ transition). Traffic breakdown is possible at any density in initial free flow that is within a density range labeled by “ $F \rightarrow S$ ”. 2D-region for synchronized flow (dashed region) is associated with the 2D-region for steady states of synchronized flow shown in Fig. 25a

about the occurrence of a growing wave of the increase in the vehicle speed that results from the $S \rightarrow F$ instability in synchronized flow is indeed supported by microscopic numerical simulations.

It should be noted that in the above qualitative explanation of the $S \rightarrow F$ instability due to the over-acceleration effect, we have not considered the speed adaptation effect. The speed adaptation effect can prevent the development of the $S \rightarrow F$ instability. The existence of the time delay in driver over-acceleration together with the speed adaptation effect that prevents the development of an $S \rightarrow F$ instability leads to the assumption that the $S \rightarrow F$ instability caused by over-acceleration should exhibit a nucleation character. The nucleation character of the $S \rightarrow F$ instability is as follows.

- A small enough local increase in the speed within synchronized flow should decay. Indeed, due to the speed adaptation effect and the time delay in over-acceleration, the small enough local increase in the speed within synchronized flow cannot cause a

growing wave of a local increase in the vehicle speed in synchronized flow leading to the development of an $S \rightarrow F$ instability. Therefore, no $S \rightarrow F$ transition occurs due to the occurrence of the small enough initial local increase in the speed within synchronized flow.

- (ii) However, if a large enough local increase in the speed within synchronized flow occurs, then the speed adaptation effect and the time delay in over-acceleration cannot prevent the growth of a wave of a local increase in the vehicle speed in synchronized flow – the $S \rightarrow F$ instability caused by over-acceleration is realized. Therefore, an $S \rightarrow F$ transition occurs due to the occurrence of the large enough initial local increase in the speed within synchronized flow.

Therefore, as already mentioned in section “Main Prediction of Three-Phase Theory,” to describe the nucleation nature of traffic breakdown, we should consider a competition between vehicle deceleration associated with the speed adaptation effect and vehicle acceleration associated with the over-acceleration effect: In the three-phase theory, the nucleation nature of traffic breakdown is explained by a competition between two opposing tendencies occurring within a random local disturbance in which the speed is lower than in an initial free flow:

- (i) A tendency towards synchronized flow due to vehicle deceleration associated with the speed adaptation effect.
- (ii) A tendency towards the initial free flow due to vehicle acceleration associated with the over-acceleration effect.

As above mentioned, the over-acceleration effect can lead to an $S \rightarrow F$ instability that causes a growing wave of a local increase of the speed in synchronized flow to the speed in free flow. The $S \rightarrow F$ instability exhibits a nucleation character. This leads to the following assumption of the three-phase theory (Kerner 2015a).

- The nucleation nature of an $S \rightarrow F$ instability at a highway bottleneck governs the nucleation nature of an $F \rightarrow S$ transition at the bottleneck. In other words, the $S \rightarrow F$ instability leads to the metastability of free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at the bottleneck.

Already in first microscopic three-phase traffic flow models, two mechanisms for over-acceleration have been introduced and studied (Kerner and Klenov 2002, 2003; Kerner et al. 2002):

- (i) Over-acceleration due to vehicle acceleration in car-following occurring without lane changing.
- (ii) Over-acceleration due to lane changing to a faster lane that is possible on a multi-lane road.

On single-lane roads, the over-acceleration mechanism occurring without lane changing does lead to the Z-characteristic for traffic breakdown ($F \rightarrow S$ transition) as well as to the increasing flow-rate function of the breakdown probability, as this has firstly been found out in numerical simulations with stochastic three-phase traffic flow models (Kerner and Klenov 2002; Kerner et al. 2002).

For multi-lane roads, the importance of the over-acceleration mechanism due to lane changing to a faster lane has been emphasized in Ref. (Kerner and Klenov 2009b). Obviously, the resulting discontinuous character of driver over-acceleration in traffic flow is determined by a combination of all possible different mechanisms of over-acceleration.

There can be developed a number of different mathematical models of driver over-acceleration leading to the discontinuously behavior of the probability of driver over-acceleration. In section “Over-Acceleration Effect,” we will discuss briefly one of them applied in the Kerner-Klenov stochastic microscopic model (Kerner and Klenov 2002, 2009b).

Basic Requirement for a Three-Phase Traffic Flow Model

Rather than a mathematical model of traffic flow, the three-phase theory is a *qualitative theory* that consists of several hypotheses (Kerner 1998a, b, 1999a, 2004a). A mathematical model of traffic flow can incorporate some of the hypotheses of the three-phase theory. It can be expected that a diverse variety of different mathematical models of traffic flow can be developed that incorporate some of the hypotheses of the three-phase theory.

It must be emphasized that not any traffic flow model that incorporate some of the hypotheses of the three-phase theory can be considered a *three-phase traffic flow model*. This is because the empirical fundamental of the three-phase theory is the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck (section “[Empirical Proof of Nucleation Nature of Traffic Breakdown \(\$F \rightarrow S\$ Transition\)](#)”). Therefore, any three-phase traffic flow model must show a Z-characteristic for traffic breakdown (Fig. 29).

In Fig. 29, we show simulations of a Z-characteristic for traffic breakdown at an on-ramp bottleneck with the Kerner-Klenov stochastic microscopic three-phase traffic flow model (Kerner and Klenov 2003). The Z-characteristic is built by metastable states of the free flow (F) and synchronized flow (S) traffic phases (solid curves F and S, respectively) together with a curve for critical nuclei for the $F \rightarrow S$ transition (a dashed curve in Fig. 29b between the metastable states of the free flow and synchronized flow phases). The Z-characteristic for traffic breakdown will be considered in more details in section “[Theoretical Z-Characteristic for Traffic Breakdown at Bottleneck](#).”

The Z-characteristic for traffic breakdown shows the flow rate range between the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$ of free flow at the bottleneck within which traffic breakdown can occur (Fig. 29b). When free flow is initially at the bottleneck, then the flow rate in this free flow downstream of the bottleneck is equal to

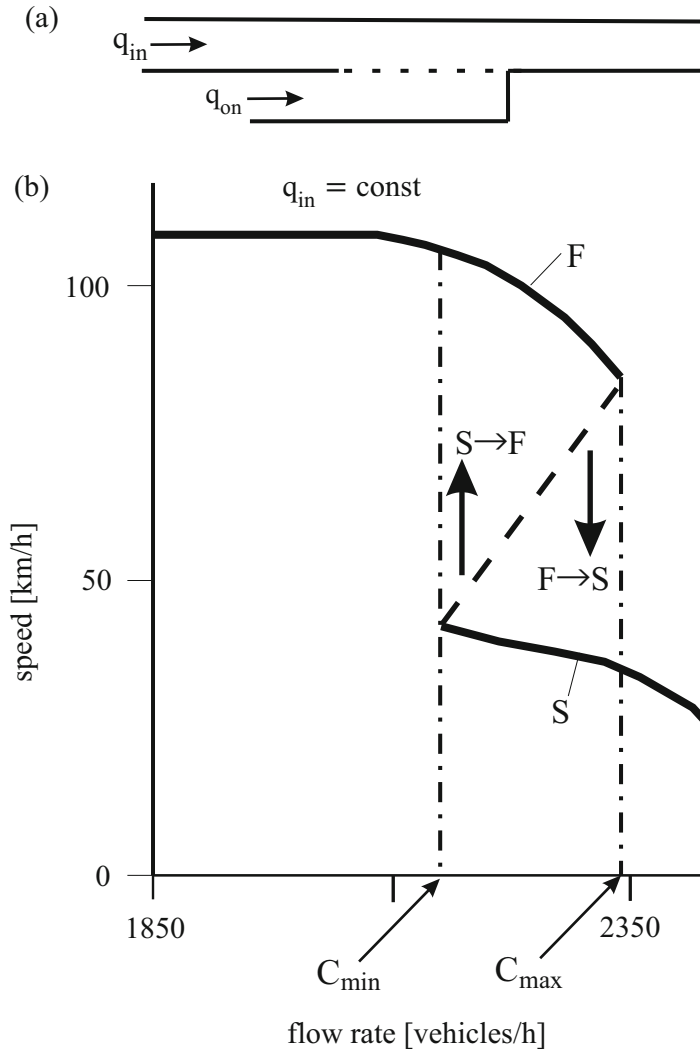
$q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$. Therefore, an infinite number of these flow rates (13) are highway capacities of free flow at the bottleneck. It should be emphasized that the Z-characteristic for traffic breakdown (Fig. 29b) is a macroscopic characteristic of traffic breakdown at the bottleneck. Some results of a microscopic theory of traffic breakdown at a bottleneck derived only recently (Kerner 2015a) will be considered in section “[S \$\rightarrow\$ F Instability at Bottleneck](#).”

The Kerner-Klenov model introduced in 2002 (Kerner and Klenov 2002) was the first mathematical three-phase traffic flow model in the framework of the three-phase theory that can show this Z-characteristic (Fig. 29b). Some months later, Kerner, Klenov, and Wolf developed a three-phase cellular automaton (CA) traffic flow model (KKW CA model) (Kerner et al. 2002). Based on the KKW CA model, the KKS (Kerner-Klenov-Schreckenberg) three-phase CA traffic flow model (Kerner et al. 2011) and the KKS_W (Kerner-Klenov-Schreckenberg-Wolf) three-phase CA traffic flow model (Kerner et al. 2013a, 2014) have been developed for a more detailed description of empirical features of real traffic within the cellular automaton approach to the three-phase theory.

In the three-phase theory, there are the free flow, synchronized flow, and wide moving jam phases. Therefore, additionally to the Z-characteristic for traffic breakdown (Fig. 29b), a three-phase traffic flow model must show a second Z-characteristic for the $S \rightarrow J$ and $J \rightarrow S$ transitions. The Z-characteristic for traffic breakdown (Fig. 29b) together with the Z-characteristic for the $S \rightarrow J$ and $J \rightarrow S$ transitions lead to a 2Z-characteristic for phase transitions (Fig. 30). The theoretical 2Z-characteristic for phase transitions (Fig. 30) is in accordance with empirical features of phase transitions in real traffic flow (Kerner 2004a) (see also Sect. 1.5 of the book (Kerner 2017a)).

These explanations lead to the following definition of the term *three-phase traffic flow model*.

- A three-phase traffic flow model is a mathematical traffic flow model in the framework of



Modeling Approaches to Traffic Breakdown, Fig. 29 Simulations of Z-characteristic for traffic breakdown with the Kerner-Klenov stochastic microscopic three-phase model: (a) Sketch of on-ramp bottleneck on a single-lane road. (b) Z-characteristic as function of the flow rate $q_{in} + q_{on}$ at the on-ramp bottleneck when the on-ramp inflow rate q_{on} is a variable and the flow rate upstream of the on-ramp bottleneck q_{in} is a given value ($q_{in} = 1850$ vehicles/h). F – free flow phase, S – synchronized flow phase. Arrow F → S illustrates symbolically

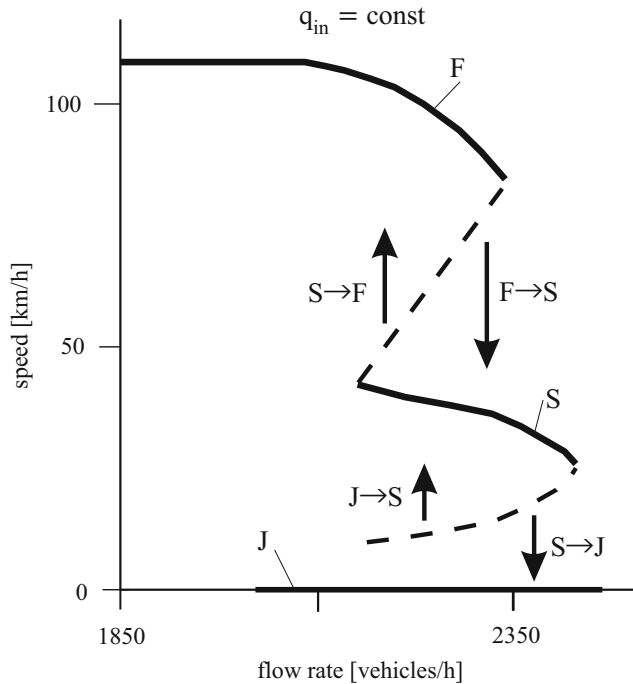
traffic breakdown (F → S transition); arrow S → F illustrates symbolically a return S → F transition. C_{min} is the minimum highway capacity, C_{max} is the maximum highway capacity. For the simplification of the consideration of the Z-characteristic, in (b) the curve S is related to an averaging of different speeds for a given q_{on} within real 2D-states of synchronized flow (S) following from simulations with the Kerner-Klenov model (Adapted from Kerner and Klenov 2003)

the three-phase theory that can show a 2Z-characteristic of phase transitions in traffic flow (Fig. 30).

- This definition means that if a traffic flow model that incorporates some of the hypotheses of the three-phase theory cannot show the

2Z-characteristic of phase transitions in traffic flow, the model cannot generally be considered a three-phase traffic flow model.

In more details, the 2Z-characteristic for phase transitions as the basic requirement for a three-



Modeling Approaches to Traffic Breakdown, Fig. 30 Simulation of 2Z-characteristic for phase transitions with the Kerner-Klenov stochastic microscopic three-phase model: 2Z-characteristic for phase transitions as function of the flow rate $q_{in} + q_{on}$ at the on-ramp bottleneck when the on-ramp inflow rate q_{on} is a variable and the flow rate upstream of the on-ramp bottleneck q_{in} is a given value ($q_{in} = 1850$ vehicles/h). The first Z-characteristic between the phases F and S is the same one as that shown in

Fig. 29b. The second Z-characteristic between the synchronized flow and wide moving jam traffic phases is discussed in the articles (Kerner 2009a, 2018) of this Encyclopedia. F – free flow phase, S – synchronized flow phase, J – wide moving jam phase. Arrows $F \rightarrow S$, $S \rightarrow F$, $S \rightarrow J$, and $J \rightarrow S$ illustrate symbolically the related phase transitions between the three traffic phase (Adapted from Kerner and Klenov 2003)

phase traffic flow model has been considered in Sect. 8.9 of the book (Kerner 2017a).

This explains the above statement that not any traffic flow model that incorporate some of the hypotheses of the three-phase theory can be considered a three-phase traffic flow model. In particular, the use of some of the hypotheses of the three-phase theory in a traffic flow model does not necessarily mean that the model becomes a three-phase traffic flow model: If this traffic flow model cannot show the 2Z-characteristic for the phase transitions, the model cannot be considered a three-phase traffic flow model.

The Kerner-Klenov stochastic three-phase traffic flow model has further been developed for different applications, in particular, to simulate on-ramp metering (see section “ANCONA On-Ramp Metering” below), speed limit control,

dynamic traffic assignment, traffic at heavy bottlenecks, traffic at moving bottlenecks, features of heterogeneous traffic flow consisting of different vehicles and drivers, features of the complexity of spatiotemporal traffic congestion observed in real measured traffic data, the diagram of congested traffic patterns at highway bottlenecks in the framework of the three-phase theory, jam warning methods, vehicle-to-vehicle (V2V) communication, the ACC (adaptive cruise control) performance of automated driving (called also automatic driving, autonomous driving, or self-driving) vehicles, traffic breakdown at signals in city traffic, over-saturated city traffic, vehicle fuel consumption in traffic networks based on a cumulative vehicle acceleration (see references to these applications of the Kerner-Klenov model in Sect. 1.12 of the book (Kerner 2017a)).

Over time several scientific groups have developed and studied new traffic flow models that incorporate some of the hypotheses of the three-phase theory: Davis (2004a) as well as Kerner and Klenov (2006c) have proposed different three-phase microscopic deterministic models; Jiang and Wu (2004), Lee et al. (2004), and Gao et al. (2007) have developed various CA models that incorporate some of the hypotheses of the three-phase theory; Siebel and Mauser (2006) and Laval (2007) have suggested macroscopic models; Jiang et al. (2007) have suggested a macroscopic version of the Kerner-Klenov speed adaptation three-phase model of Ref. (Kerner and Klenov 2006c). Some simulation results of traffic flow models incorporating hypotheses of the three-phase theory can be found, for example, in Ref. (Davis 2004b, 2006a, b, 2007, 2008, 2014, 2016; He et al. 2010; Hu et al. 2012; Jiang et al. 2015, 2017; Jiang and Wu 2005, 2007; Kokubo et al. 2011; Li et al. 2007; Qian et al. 2017; Pottmeier et al. 2007; Tian et al. 2016a, b; Tian et al. 2015, 2016c; Wang et al. 2007; Xiang et al. 2013; Yang et al. 2013) (see also references in Sect. 1.12 of the book (Kerner 2017a)).

There are a number of mathematical formulations of the hypotheses of the three-phase theory (see, e.g., (Davis 2004b, 2006a, b, 2007, 2008, 2014, 2016; Gao et al. 2007; He et al. 2010; Hu et al. 2012; Jiang et al. 2015; Jiang and Wu 2004, 2005, 2007; Jiang et al. 2014; Kerner and Klenov 2002, 2003, 2006c; Kerner et al. 2002, 2006a, b, 2007, 2011; Kerner et al. 2013a, 2014; Kokubo et al. 2011; Lee et al. 2004; Li et al. 2007; Qian et al. 2017; Pottmeier et al. 2007; Tian et al. 2016a, b, c; Tian et al. 2015; Wang et al. 2007; Xiang et al. 2013; Yang et al. 2013)). It is not possible to consider all these different mathematical formulations in the review article. Therefore, we consider only a discrete version (Kerner and Klenov 2009b) of the stochastic microscopic model of Kerner and Klenov (2002, 2003). Moreover, we limit the consideration by an analysis of driver behavioral assumptions of the three-phase theory incorporated in this model for a single-lane road. This limitation of the model analysis permits us to explain the hypothesis of the three-phase theory of the nucleation nature of traffic

breakdown ($F \rightarrow S$ transition) at a highway bottleneck (a detailed formulation of the Kerner-Klenov microscopic stochastic three-phase traffic flow model for multi-lane roads with different types of bottlenecks can be found in Appendix A of the book (Kerner 2017a)).

Kerner-Klenov Stochastic Microscopic Three-Phase Traffic Flow Model

Basic Rules of Vehicle Motion

In a discrete model version of the Kerner-Klenov stochastic microscopic three-phase model considered here, rather than the continuum space coordinate (Kerner and Klenov 2002, 2003), a discretized space coordinate with a small enough value of the discretization space interval δx is used (Kerner and Klenov 2009b). Consequently, the vehicle speed and acceleration (deceleration) discretization intervals are $\delta v = \delta x / \tau$ and $\delta a = \delta v / \tau$, respectively, where τ is time step. Because in the discrete model version discrete (and dimensionless) values of space coordinate, speed and acceleration are used, which are measured, respectively, in values δx , δv , and δa , and time is measured in values of τ , value τ in all formulas is assumed below to be the dimensionless value $\tau = 1$. A choice of $\delta x = 0.01$ m made in the model determines the accuracy of vehicle speed calculations *in comparison* with the initial continuum in space stochastic model of (Kerner and Klenov 2002, 2003). We have found that the discrete model exhibits similar characteristics of phase transitions and resulting congested patterns at highway bottlenecks as those in the continuum model at δx that satisfies the conditions

$$\delta x / \tau^2 \ll a, a^{(a)}, a^{(b)}, a^{(0)}, \quad (85)$$

where model parameters for driver acceleration a , $a^{(a)}$, $a^{(b)}$, $a^{(0)}$ will be explained below.

Update rules of vehicle motion in the discrete model for identical drivers and identical vehicles moving in a road lane are as follows (Kerner and Klenov 2003, 2009b):

$$v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}_{n+1} + \xi_n, v_n + a\tau, v_{s,n})), \quad (86)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (87)$$

where the index n corresponds to the discrete time $t_n = \tau n$, $n = 0, 1, \dots$; v_n is the vehicle speed at time step n , a is the maximum acceleration, $\tilde{v}_n$ is the vehicle speed without speed fluctuations ξ_n :

$$\tilde{v}_{n+1} = \min(v_{\text{free}}, v_{s,n}, v_{c,n}), \quad (88)$$

$$v_{c,n} = \begin{cases} v_n + \Delta_n & \text{at } g_n \leq G_n \\ v_n + a_n\tau & \text{at } g_n > G_n, \end{cases} \quad (89)$$

$$\Delta_n = \max(-b_n\tau, \min(a_n\tau, v_{\ell,n} - v_n)), \quad (90)$$

$$g_n = x_{\ell,n} - x_n - d, \quad (91)$$

the subscript ℓ marks variables related to the preceding vehicle, $v_{s,n}$ is a safe speed at time step n , v_{free} is the free flow speed, ξ_n describes speed fluctuations; g_n is a space gap between two vehicles following each other; G_n is the synchronization space gap; all vehicles have the same length d . The vehicle length d includes the mean space gap between vehicles that are in a standstill within a wide moving jam. Values $a_n \geq 0$ and $b_n \geq 0$ in Eqs. 89 and 90 restrict changes in speed per time step when the vehicle accelerates or adjusts the speed to that of the preceding vehicle.

Equations 89 and 90 describe the adaptation of the vehicle speed to the speed of the preceding vehicle, i.e., the speed adaptation effect in synchronized flow (section “[Competition of Driver Speed Adaptation with Driver Over-Acceleration in Three-Phase Theory](#)”). This vehicle speed adaptation takes place within the synchronization gap G_n : At

$$g_n \leq G_n \quad (92)$$

the vehicle tends to adjust its speed to the speed of the preceding vehicle. This means that the vehicle decelerates if $v_n > v_{\ell,n}$, and accelerates if $v_n < v_{\ell,n}$.

In Eq. 89, the synchronization gap G_n depends on the vehicle speed v_n and on the speed of the preceding vehicle $v_{\ell,n}$:

$$G_n = G(v_n, v_{\ell,n}), \quad (93)$$

$$G(u, w) = \max(0, \lfloor k\tau u + a^{-1}u(u - w) \rfloor), \quad (94)$$

where $k > 1$ is constant; $\lfloor z \rfloor$ denotes the integer part of z .

The speed adaptation effect within the synchronization distance is related to the hypothesis of the three-phase theory: Hypothetical steady states of synchronized flow cover a 2D region in the flow–density (see Fig. 31a). Boundaries F , L , and U of this 2D-region shown in Fig. 31a are, respectively, associated with the free flow speed, a synchronization space gap G , and a safe space gap g_{safe} . A speed-function of the safe space gap $g_{\text{safe}}(v)$ in steady states is found from the equation

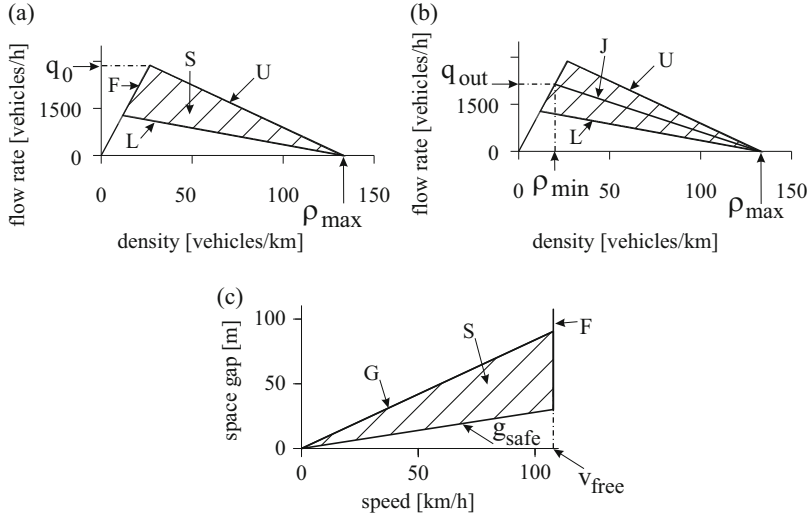
$$v = v_s(g_{\text{safe}}, v). \quad (95)$$

Respectively, as for the continuum model (see Sect. 16.3 of the book (Kerner 2004a)), for the discrete model hypothetical steady states of synchronized flow cover a 2D-region in the flow–density plane (Fig. 31a, b). However, because the speed v and space gap g are integer in the discrete model, the steady states do not form a continuum in the flow–density plane as they do in the continuum model. The inequalities

$$v \leq v_{\text{free}}, g \leq G(v, v), g \geq g_{\text{safe}}(v), \quad (96)$$

define a 2D-region in the space-gap–speed plane (Fig. 31c) in which the hypothetical steady states exist for the discrete model, when all model fluctuations are neglected.

In Eq. 96, we have taken into account that in the hypothetical steady states of synchronized flow vehicle speeds and space gaps are assumed to be time-independent and the speed of each of the vehicles is equal to the speed of the associated preceding vehicle: $v = v_{\ell}$. However, due to model fluctuations, steady states of synchronized flow are destroyed, i.e., they do not exist in



Modeling Approaches to Traffic Breakdown, Fig. 31 Steady speed states for the Kerner-Klenov traffic flow model in the flow-density (a, b) and in the space-gap-speed planes (c). In (a, b), L and U are, respectively, lower and upper boundaries of 2D-regions of steady states of synchronized flow. In (b), J is the line J whose slope is

equal to the characteristic mean velocity v_g of a wide moving jam; in the flow-density plane, the line J represents the propagation of the downstream front of the wide moving jam with time-independent mean velocity v_g . F free flow, S synchronized flow

simulations; this explains the term “hypothetical” steady states of synchronized flow. Therefore, rather than steady states some nonhomogeneous in space and time traffic states occur. In other words, steady states do not realized in real simulations. Driver time delays are described through model fluctuations. Therefore, any application of the Kerner-Klenov stochastic microscopic three-phase traffic flow model without model fluctuations has *no sense*. In other words, for the description of real spatiotemporal traffic flow phenomena, model speed fluctuations incorporated in this model are needed.

In the model, random vehicle deceleration and acceleration are applied depending on whether the vehicle decelerates or accelerates, or else maintains its speed:

$$\xi_n = \begin{cases} \xi_a & \text{if } S_{n+1} = 1 \\ -\xi_b & \text{if } S_{n+1} = -1 \\ \xi^{(0)} & \text{if } S_{n+1} = 0. \end{cases} \quad (97)$$

State of vehicle motion S_{n+1} in Eq. 97 is determined by formula

$$S_{n+1} = \begin{cases} -1 & \text{if } \tilde{v}_{n+1} < v_n \\ 1 & \text{if } \tilde{v}_{n+1} > v_n \\ 0 & \text{if } \tilde{v}_{n+1} = v_n. \end{cases} \quad (98)$$

In Eq. 97, ξ_b , $\xi^{(0)}$, and ξ_a are random sources for deceleration and acceleration that are as follows:

$$\xi_b = a^{(b)} \tau \Theta(p_b - r), \quad (99)$$

$$\xi^{(0)} = a^{(0)} \begin{cases} -1 & \text{if } r < p^{(0)} \\ 1 & \text{if } p^{(0)} \leq r < 2p^{(0)} \text{ and } v_n > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (100)$$

$$\xi_a = a^{(a)} \tau \Theta(p_a - r), \quad (101)$$

p_b is probability of random vehicle deceleration, p_a is probability of random vehicle acceleration, $p^{(0)}$ and $a^{(0)} \leq a$ are constants, $r = \text{rand}(0, 1)$, $\Theta(z) = 0$ at $z < 0$ and $\Theta(z) = 1$ at $z \geq 0$, $a^{(a)}$ and $a^{(b)}$ are model parameters, which in some

applications can be chosen as speed functions $a^{(a)} = a^{(a)}(v_n)$ and $a^{(b)} = a^{(b)}(v_n)$ (see model parameters in Appendix A of the book (Kerner 2017a)).

To simulate time delays either in vehicle acceleration or in vehicle deceleration, a_n and b_n in Eq. 90 are taken as the following stochastic functions

$$a_n = a\Theta(P_0 - r_1), \quad (102)$$

$$b_n = a\Theta(P_1 - r_1), \quad (103)$$

$$P_0 = \begin{cases} p_0 & \text{if } S_n \neq 1 \\ 1 & \text{if } S_n = 1, \end{cases} \quad (104)$$

$$P_1 = \begin{cases} p_1 & \text{if } S_n \neq -1 \\ p_2 & \text{if } S_n = -1, \end{cases} \quad (105)$$

$r_1 = \text{rand}(0, 1)$, p_1 is constant, $p_0 = p_0(v_n)$ and $p_2 = p_2(v_n)$ are speed functions.

In the model, the safe speed $v_{s,n}$ in (86) is chosen in accordance with Gipps's equation for the safe speed (Gipps 1981) and further developments of the Gipps's equation made by Krauß et al. (1997) (see for details section A.3.5 of the book (Kerner 2017a)).

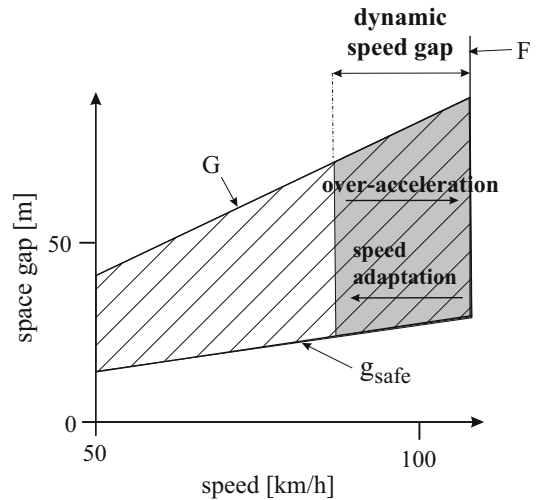
Over-Acceleration Effect

In the stochastic model, there are several different mathematical approaches for the simulation of the over-acceleration effect (see section A.8 of the book (Kerner 2017a)). In each of these approaches, a *dynamic speed gap* appears (gray region labeled by “dynamic speed gap” in Fig. 32) between free flow (line F) and steady states of synchronized flow of lower speeds. Within this dynamic speed gap, initially steady states of synchronized flow of higher speeds are destroyed and no long-time living synchronized flow states occur.

In the stochastic model, a *discontinuous character* of over-acceleration discussed in section “Driver Over-Acceleration” is modeled through the occurrence of the dynamic speed gap caused by stochastic model dynamics that destroys initially steady states of synchronized flow of higher speeds (gray region in Fig. 32). A competition

between the speed adaptation and over-acceleration effects (labeled by “speed adaptation” and “over-acceleration” in Fig. 32), which occurs within a local disturbance in free flow, is associated with this dynamic speed gap.

In the simple version of the Kerner-Klenov model discussed here (section “Basic Rules of Vehicle Motion”), the dynamic speed gap between states of free flow and synchronized flow resulting in a discontinuous character of over-acceleration can be simulated through the following stochastic approach. The function ξ_a in Eq. 97 is given by formula (101). In accordance with Eq. 97, these model fluctuations (101) are applied only if the vehicle should accelerate without fluctuations. The probability of the driver over-acceleration p_a in Eq. 101 is less than 1. This simulates a time delay of the driver over-acceleration required for the occurrence of an $S \rightarrow F$ instability in synchronized flow (section “S→F Instability at Bottleneck”).



Modeling Approaches to Traffic Breakdown,

Fig. 32 Qualitative explanation of competition of over-acceleration and speed adaptation effects in the Kerner-Klenov stochastic microscopic model. Parts of steady states of synchronized flow at higher speeds (dashed region) and free flow speed (F). Gray region shows qualitatively a dynamic speed gap in which steady states of synchronized flow are destroyed and no long-time living synchronized flow states occur (Adapted from Kerner 2009c)

Speed Adaptation Effect

We consider a vehicle speed v_n that satisfies conditions

$$0 < v_n < v_{\text{free}}. \quad (106)$$

Within a 2D-region of states of synchronized flow (2D-dashed region in Fig. 33) that satisfy conditions

$$g_{\text{safe},n} < g_n \leq G_n, \quad (107)$$

we assume that conditions

$$\begin{aligned} |\Delta v_n| &\leq a\tau, a_n = b_n = a, v_n + \Delta_n + \xi_n \\ &\leq \min(v_{\text{free}}, v_{s,n}, v_n + a\tau), \end{aligned} \quad (108)$$

are satisfied. In Eq. 108, Δv_n is the speed difference between the speed of the preceding vehicle $v_{\ell,n}$ and the vehicle speed v_n :

$$\Delta v_n = v_{\ell,n} - v_n. \quad (109)$$

Under conditions (107) and (108), from Eqs. 86, 88, 89, and 90 with some probability the vehicle speed v_{n+1} at the next time step $n+1$ is equal to

$$v_{n+1} = v_n + \Delta v_n + \xi_n. \quad (110)$$

Note that the vehicle acceleration (deceleration) is equal to

$$A_n = (v_{n+1} - v_n)/\tau. \quad (111)$$

This means that accordingly to Eq. 110 the vehicle acceleration (deceleration) is equal to

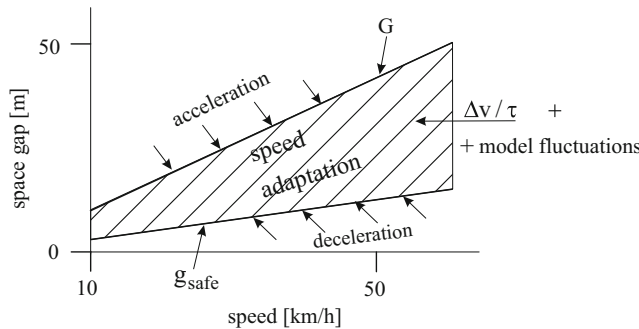
$$A_n = (\Delta v_n + \xi_n)/\tau. \quad (112)$$

Equation 112 that is valid under conditions (107) is a discrete version of Eq. 83 discussed in section “Speed Adaptation Effect Within Dynamic 2D-States of Synchronized Flow.” Equation 112 explains the speed adaptation effect in the stochastic model (labeled “speed adaptation” in Fig. 33): Within the space gap range (107) the speed difference Δv_n (109) in Eq. 112 describes speed adaptation to the speed of the preceding vehicle that occurs without caring what the precise space gap to the preceding vehicle is. In addition to this deterministic effect, in the stochastic model there is a random speed change ξ_n , which leads to a stochastic behavior of speed adaptation in the model.

Competition Between Speed Adaptation with Over-Acceleration

In accordance with Eq. 97, the stochastic behavior of speed adaptation in the model (112) can be rewritten as follows

$$A_n = \Delta v_n/\tau + \begin{cases} \xi_a/\tau & \text{if } S_{n+1} = 1 \\ -\xi_b/\tau & \text{if } S_{n+1} = -1 \\ \xi^{(0)}/\tau & \text{if } S_{n+1} = 0, \end{cases} \quad (113)$$



Modeling Approaches to Traffic Breakdown, Fig. 33 Explanation of the speed adaptation effect within 2D-region of states of synchronized flow. A part of steady states of synchronized flow of lower speeds associated with

Eq. 106 taken from Fig. 31c. $\Delta v = v_{\ell} - v$ is difference between the speed of the preceding vehicle v_{ℓ} and the vehicle speed v

where the state of the vehicle motion S_{n+1} is given by Eq. 98.

If at time step n the speed difference $\Delta v_n > 0$, then in accordance with the speed adaption effect a vehicle accelerates, i.e., $S_{n+1} = 1$. In this case, from Eq. 113 we get

$$A_n = \Delta v_n / \tau + \xi_a / \tau \quad (114)$$

However, in accordance with Eq. 101, model fluctuations ξ_a describe the over-acceleration effect. Thus, we see that the vehicle acceleration (deceleration) A_n in Eq. 114 is the sum of the deterministic vehicle speed adaptation described by the term $\Delta v_n / \tau$ and a stochastic description of over-acceleration effect described by fluctuations in vehicle acceleration, i.e., by the term ξ_a / τ . This means that under condition $S_{n+1} = 1$ within 2D-states of synchronized flow model rather than the “pure” speed adaptation described by the term $\Delta v_n / \tau$ there is always a competition

between the speed adaptation effect and the over-acceleration effect as stated in the three-phase theory. If the over-acceleration overcomes the speed adaptation, an $S \rightarrow F$ instability occurs. The uninterrupted development of the $S \rightarrow F$ instability leads to a transition from synchronized flow to free flow ($S \rightarrow F$ transition). Otherwise, the speed adaptation effect maintains synchronized flow.

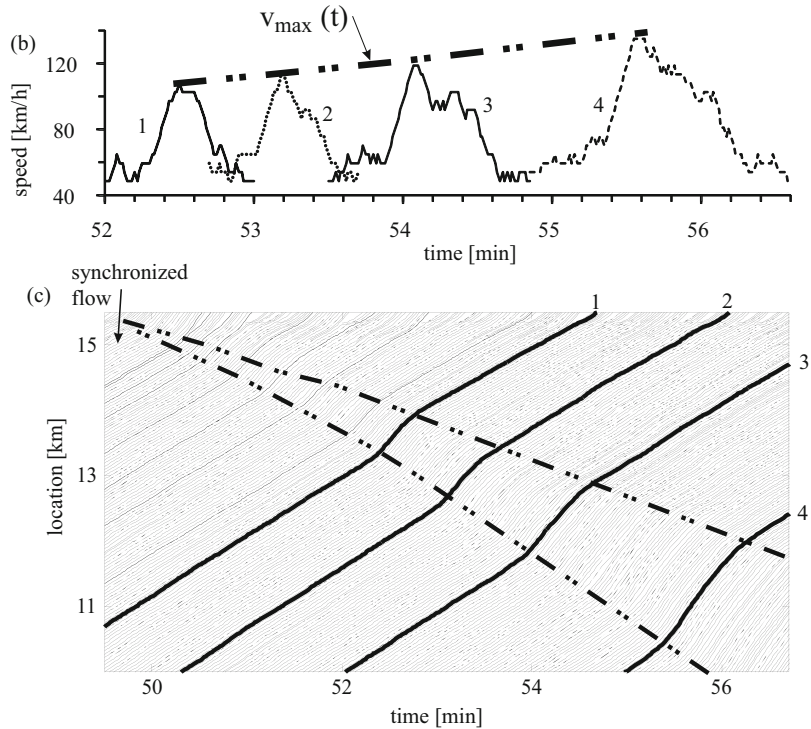
$S \rightarrow F$ Instability at Bottleneck

In accordance with the conclusion of section “Competition Between Speed Adaptation with Over-Acceleration,” when in a local speed increase within synchronized flow the over-acceleration overcomes the speed adaptation, an $S \rightarrow F$ instability occurs. The uninterrupted development of the $S \rightarrow F$ instability leads to a transition from synchronized flow to free flow ($S \rightarrow F$ transition) (Figs. 34 and 35). The $S \rightarrow F$ instability

Modeling Approaches to Traffic Breakdown,

Fig. 34 $S \rightarrow F$ instability introduced in three-phase theory (Kerner 2015a): Growing wave of speed increase in synchronized flow due to driver over-acceleration as time-function (a) and on vehicle trajectories (b). Dashed-dotted curve in (a) marks the development of the maximum vehicle speed $v_{\max}$ within a local speed increase in synchronized flow. Dashed-dotted curves in (b) mark the development of the local speed increase in synchronized flow in space and time. Numbers 1–4 mark the same vehicles in (a) and (b)

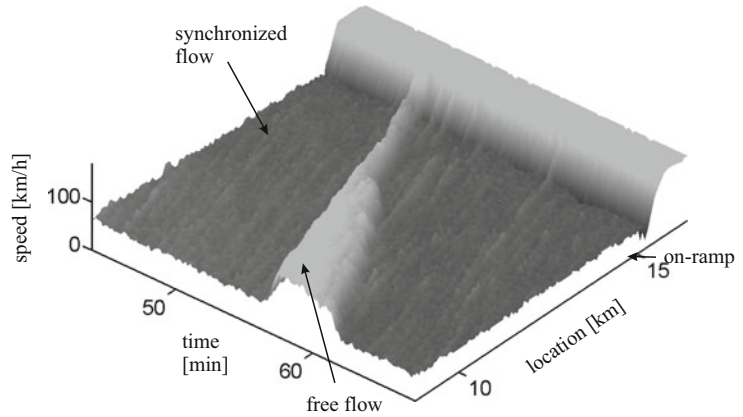
The $S \rightarrow F$ instability introduced in three-phase theory: Growing wave of speed **increase** in synchronized flow due to driver over-acceleration



Modeling Approaches to Traffic Breakdown,

Fig. 35 Continuation of Fig. 34. Subsequent development of the $S \rightarrow F$ instability (Kerner 2015a): The $S \rightarrow F$ transition is the emergence of free flow from the growing wave of speed increase in synchronized flow due to driver over-acceleration

The $S \rightarrow F$ instability introduced in three-phase theory: Growing wave of speed **increase** in synchronized flow due to driver over-acceleration leading to the transition from synchronized flow to free flow



is a growing wave of a local **increase** in the speed in synchronized flow (Fig. 34a, b).

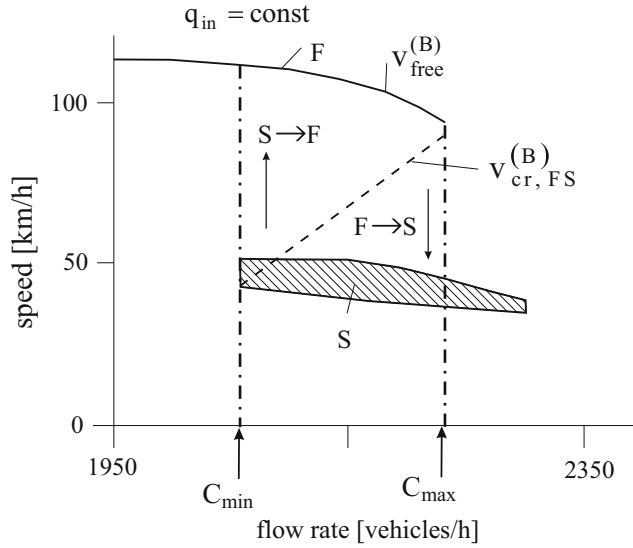
In a microscopic theory of an $S \rightarrow F$ instability at a highway bottleneck (Kerner 2015a), it has been found that the $S \rightarrow F$ instability governs the occurrence of an $F \rightarrow S$ transition at a highway bottleneck. Additionally, it has been found that the $S \rightarrow F$ instability exhibits the nucleation nature. In its turn, the nucleation nature of the $S \rightarrow F$ instability determines the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. In particular, after an $F \rightarrow S$ transition has begun to develop at the bottleneck, an $S \rightarrow F$ instability can occur spontaneously within the emergent synchronized flow. Due to this $S \rightarrow F$ instability, an $S \rightarrow F$ transition occurs: synchronized flow dissolves and free flow recovers at the bottleneck. The $F \rightarrow S$ transition together with the subsequent $S \rightarrow F$ transition has been called a sequence of $F \rightarrow S \rightarrow F$ transitions (Kerner 2015a). In (Kerner 2015a) it has been shown that sequences of $F \rightarrow S \rightarrow F$ transitions are responsible for a random time delay to traffic breakdown at the bottleneck (see for more detail explanations section 5.13 of the book (Kerner 2017a)). Recently, empirical sequences of $F \rightarrow S \rightarrow F$ transitions appearing before traffic breakdown have occurred at a bottleneck have been found in microscopic probe vehicle data (Molzahn et al. 2017).

Thus, the nucleation nature of the $S \rightarrow F$ instability explains the empirical fundamental of transportation science – the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck (section “[Empirical Proof of Nucleation Nature of Traffic Breakdown \(\$F \rightarrow S\$ Transition\)](#)”).

Theoretical Z-Characteristic for Traffic Breakdown at Bottleneck

Microscopic stochastic three-phase theory of traffic breakdown ($F \rightarrow S$ transition) has been in detailed presented in chapter “[Freeway Traffic Management and Control](#)” of the book (Kerner 2017a). Here, based on this theory, we discuss macroscopic features of the breakdown with the use of the Z-characteristic for traffic breakdown at the bottleneck that has been briefly discussed in section “[Empirical Nucleation Nature of Traffic Breakdown as Origin of The Infinity of Highway Capacities at Any Time Instant](#)” (Kerner and Klenov 2003).

As a qualitative Z-characteristic for traffic breakdown shown in Fig. 16, the theoretical Z-characteristic for traffic breakdown (Fig. 36) is built by metastable states of the free flow (F) and synchronized flow (S) traffic phases (solid curve F and hatched 2D-region S, respectively) together with a branch for the speed within critical nuclei



Modeling Approaches to Traffic Breakdown, Fig. 36 Simulations of Z-characteristic for traffic breakdown at on-ramp bottleneck with the Kerner-Klenov stochastic microscopic three-phase model: Z-characteristic as function of the flow rate at the on-ramp bottleneck $q_{in} + q_{on}$ when the on-ramp inflow rate q_{on} is a variable and the flow

rate upstream of the on-ramp bottleneck q_{in} is a given value ($q_{in} = 1756$ vehicles/h). F free flow phase, S synchronized flow phase. Arrows $F \rightarrow S$ and $S \rightarrow F$ illustrate qualitatively the $F \rightarrow S$ and $S \rightarrow F$ transitions, respectively. C_{min} is the minimum highway capacity, C_{max} is the maximum highway capacity (Adapted from Kerner and Klenov 2003)

$v_{cr,FS}^{(B)}(q_{sum})$ for the $F \rightarrow S$ transition (traffic breakdown) (dashed curve $v_{cr,FS}^{(B)}$ in Fig. 36 between the metastable states of the free flow and synchronized flow traffic phases).

The theoretical Z-characteristic for traffic breakdown (Fig. 36) derived through simulations of the Kerner-Klenov microscopic stochastic three-phase model exhibits the following features.

- (i) When an initial state of traffic flow at the bottleneck is free flow, simulations show that if the speed within a local speed disturbance in a free flow state at the bottleneck is equal to or smaller than the speed $v_{cr,FS}^{(B)}$ within a critical nucleus for traffic breakdown, this disturbance is a nucleus for an $F \rightarrow S$ transition: The $F \rightarrow S$ transition (traffic breakdown) does occur at the bottleneck.
- (ii) At the flow rate $q_{sum} < C_{min}$, there is no nucleus that can lead to an $F \rightarrow S$ transition at the bottleneck.

- (iii) The speed $v_{cr,FS}^{(B)}$ within the critical nucleus for traffic breakdown is the smallest one at the flow rate $q_{sum} = C_{min}$.
- (iv) Within the flow rate range $C_{min} \leq q_{sum} < C_{max}$ (13), when the flow rate q_{sum} increases, the amplitude of the critical nucleus $\Delta v_{cr,FS}^{(B)} = v_{free}^{(B)} - v_{cr,FS}^{(B)}$ required for the $F \rightarrow S$ transition at the bottleneck decreases (here $v_{free}^{(B)}$ is the averaged free flow speed at the bottleneck (Fig. 36)).
- (v) This decrease in the amplitude of the critical nucleus for traffic breakdown $v_{cr,FS}^{(B)}$ with the increase in the flow rate q_{sum} found in simulations explains another result of simulations that the probability of traffic breakdown $P^{(B)}(q_{sum})$ during a given time interval T_{ob} is an increasing function of the flow rate q_{sum} .

It should be emphasized that as the qualitative Z-characteristic shown in Fig. 16, the theoretical

Z-characteristic for traffic breakdown (Fig. 36) is a macroscopic characteristic of traffic breakdown. This means that it has been derived based on averaging of many simulation realizations (runs) in which traffic breakdown has been observed in simulations. These different simulation realizations (runs) have been made for each of the flow rates q_{sum} . For this reason, the branch for the speed within the critical nuclei required for the breakdown (dashed curve in Fig. 36) represents averaged values of the speed within the critical nuclei. In the reality, as found in empirical data (Kerner et al. 2015) and in the microscopic theory of the breakdown nucleation (Kerner 2015a), a nucleus for traffic breakdown exhibits a very complex spatiotemporal microscopic structure.

Physics of Description of Stochastic Driver Acceleration and Deceleration

For the deeper understanding of the Kerner-Klenov model, in this section we discuss the motivation for the description of stochastic driver acceleration and deceleration made in the model. We have accepted the following driver behavioral assumptions:

- (i) In the same driving situation, in which a driver should accelerate, the driver can randomly apply different values of acceleration. In other words, the choice of the value of driver acceleration is a stochastic process.
- (ii) Respectively, we have assumed that in the same driving situation, in which a driver should decelerate, the driver can apply different random values of deceleration. Thus, the choice of the value of driver deceleration is also a stochastic process.
- (iii) Both stochastic driver acceleration and stochastic driver deceleration occur with some driver time delays that are random values. The mean value of each of the driver time delays can considerably depend on the driving situation.

To distinguish between cases, in which a driver should accelerate or the driver should decelerate, the state of the vehicle motion S_{n+1} (98) has been introduced in the model:

$$S_{n+1} = 1 \quad (115)$$

denotes a case in which the driver should accelerate, whereas

$$S_{n+1} = -1 \quad (116)$$

denotes a case in which the driver should decelerate.

Stochastic Driver Acceleration and Its Time Delay
Under condition (115), due to a value ξ_a of random driver acceleration, the vehicle speed at the next time step $n + 1$ can randomly increase additionally to the speed $\tilde{v}_{n+1}$ (88) calculated in accordance with values of the safe speed $v_{s,n}$, the space gap g_n , and the speed difference $v_{\ell,n} - v_n$ between the preceding vehicle and the vehicle as formulated in Eqs. 89 and 90.

In its turn, to simulate a random driver time delay in acceleration, in Eqs. 89 and 90 the value of acceleration a_n is taken as a stochastic function (102). A random driver time delay in acceleration occurs *only*, when at the previous time step n the vehicle does not still accelerate ($S_n \neq 1$). In this case, the mean value of the random driver time delay in acceleration is determined by the choice of probability p_0 in (104). Accordingly to (104), for a case when at the previous time step n the vehicle accelerates already ($S_n = 1$) the value of probability P_0 in (102) is chosen 1; this means that in the model there is no random interruption of driver acceleration.

In any case, the maximum driver acceleration is limited by value a . Therefore, the maximum increase in the speed at one step caused by a stochastic value of driver acceleration is limited by $v_{n+1} = v_n + a$.

Stochastic Driver Deceleration and its Time Delay
Under condition (116), due to random deceleration value $-\xi_b$, the vehicle speed v_{n+1} at the next time step $n + 1$ can randomly decrease additionally to the speed $\tilde{v}_{n+1}$ (88) calculated in accordance with values of the safe speed $v_{s,n}$, the space gap g_n , and the speed difference $v_{\ell,n} - v_n$ between the speed of the preceding vehicle and the vehicle speed, as formulated in Eqs. 89 and 90.

In its turn, to simulate a random driver time delay in deceleration, in Eq. 90 the value of driver deceleration b_n is taken as a stochastic function (103). The time delay in deceleration is related to a case, in which at the previous time step n the vehicle does not still decelerate ($S_n \neq -1$). In accordance with Eq. 105, the mean value of the time delay in deceleration is determined by the choice of probability p_1 .

Accordingly to Eq. 105, for a case when at the previous time step n the vehicle decelerates already ($S_n = -1$) probability P_1 in Eq. 103 is equal to probability p_2 that can be chosen less as 1. In this case, with probability $1 - p_2$ the random driver deceleration $b_n = 0$. Therefore, as follows from Eqs. 89 and 90, the vehicle that has decelerated at time step n interrupts the deceleration even if the vehicle moves faster than the preceding vehicle ($v_{\ell,n} - v_n < 0$). It should be noted that in accordance with formulas (89) and (90) this random interruption of driver deceleration can occur only if the space gap of the vehicle satisfies condition (107), i.e., when the space gap is smaller than the synchronization space gap and it is larger than the safe space gap. In this case, due to the interruption of vehicle deceleration, the vehicles come closer to the preceding vehicle within the synchronization space gap. The random interruption of vehicle deceleration is used for simulation of moving jam emergence in synchronized flow ($S \rightarrow J$ instability). A consideration of the $S \rightarrow J$ instability is out of scope of this article; it is considered in the article (Kerner 2018) of this Encyclopedia.

Nearly Zero Driver Acceleration While Moving at Time-Independent Speed

There are driving situations in which a driver has to follow the preceding vehicle that moves at a time-independent speed and the driver cannot pass it. Such situations occur often in city traffic when the overtaken is prohibited and a road is a single-lane road (in the direction of traffic). However, during a long enough time interval, a driver cannot maintain the acceleration (deceleration) of vehicle to be exactly equal to zero. Indeed, at some random time instants (usually with intervals

of several seconds), the driver has to make small corrections in vehicle acceleration. This means that the vehicle speed can be maintained by the driver only on average time-independent. Due to the random small corrections in vehicle acceleration, the driver acceleration can be considered a stochastic time-function that mean amplitude is a very small one in comparison with usual acceleration values.

To simulate such stochastic driver acceleration, the following development has been made in the model (Kerner and Klenov 2009b). When the vehicle speed should be maintained to be time-independent, specifically, when from Eq. 88 it follows that

$$\tilde{v}_{n+1} = v_n, \quad (117)$$

for the state of the vehicle motion S_{n+1} in Eq. 98 we get

$$S_{n+1} = 0. \quad (118)$$

Under condition (118), in accordance with Eq. 97, due to random value $\xi^{(0)}$ that can be randomly a positive or negative one, the vehicle speed v_{n+1} at the next time step $n+1$ can randomly increase (acceleration) or decrease (deceleration) additionally to the speed $\tilde{v}_{n+1}$ (88). In formula (100) that describes the value $\xi^{(0)}$, for simulations we have used a small enough value $\xi^{(0)}$ (in comparison with the maximum acceleration). Moreover, due to the choice of small probability of the occurrence of this stochastic driver acceleration or deceleration, the mean time interval between random events of the stochastic driver acceleration or deceleration is considerably longer than time step (1 s) used in the model. This model feature should simulate both long time intervals between corrections in vehicle acceleration or deceleration occurring when a driver should on average maintain a time-independent speed and that the mean amplitude of these corrections is a small one in comparison with usual vehicle acceleration or deceleration values.

Competition between Driver Speed Adaptation with Driver Over- Deceleration: Origin of $S \rightarrow J$ Instability

Let us return to a consideration of formula (113). If at time step n the speed difference $\Delta v_n < 0$, then in accordance with the speed adaptation effect a vehicle decelerates, i.e., $S_{n+1} = -1$. In this case, from Eq. 113 we get formula

$$A_n = \Delta v_n / \tau - \xi_b / \tau. \quad (119)$$

in which, as above mentioned, $-\xi_b$ is a random source for vehicle deceleration described by formula (99). As emphasized in section 16.3.7 of the book (Kerner 2004a), formula (99) describes the over-deceleration effect (driver over reaction).

Thus, we see that the vehicle acceleration (deceleration) A_n in Eq. 119 is the sum of the deterministic vehicle speed adaptation described by the term $\Delta v_n / \tau$ and a stochastic description of over-deceleration effect described by fluctuations in vehicle deceleration, i.e., by the term $-\xi_b / \tau$. This means that under condition $S_{n+1} = -1$ within 2D-states of synchronized flow rather than the “pure” speed adaptation described by the term $\Delta v_n / \tau$ there is always a competition between the speed adaptation effect and the over-deceleration effect as stated in the three-phase theory. If the over-deceleration overcomes the speed adaptation, an $S \rightarrow J$ instability occurs. The uninterrupted development of the $S \rightarrow J$ instability leads to a transition from synchronized flow to a wide moving jam ($S \rightarrow J$ transition). Otherwise, the speed adaptation effect maintains synchronized flow.

The $S \rightarrow J$ instability results from a competition between the driver speed adaptation and the driver over-deceleration effect caused by the driver reaction time. As explained in section “Classical General Motors (GM) Model Class: Free Flow Instability due to Driver Reaction Time,” the driver over-deceleration effect is the reason for the classical traffic flow instability of Herman, Montroll, Potts, Rothery, Chandler, Gazis (Chandler et al. 1958; Gazis et al. 1959, 1961; Herman et al. 1959). Therefore, the $S \rightarrow J$ instability is associated with the classical traffic flow instability (Kerner 2004a, 2009c, 2017a).

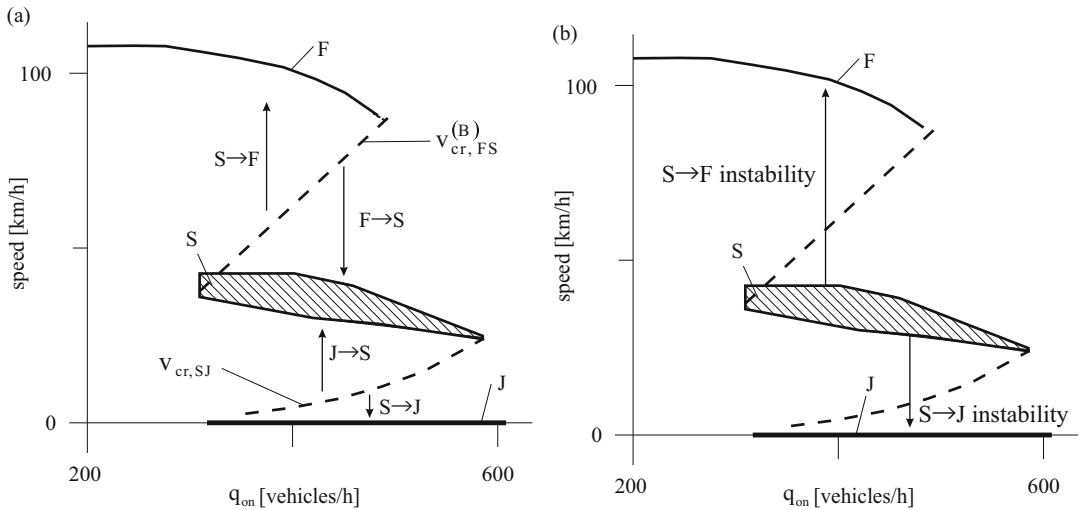
In more details, the $S \rightarrow J$ instability will be considered in the article (Kerner 2018) of this Encyclopedia.

Theoretical Double Z-Characteristic for Phase Transitions and Two Instabilities in Synchronized Flow

As mentioned in section “Theoretical Z-Characteristic for Traffic Breakdown at Bottleneck,” the metastability of free flow at a highway bottleneck with respect to an $F \rightarrow S$ transition is described by the Z-characteristic for traffic breakdown (Fig. 36). Synchronized flow related to this Z-characteristic is metastable with respect to a return $S \rightarrow F$ transition. The synchronized flow is also metastable with respect to an $S \rightarrow J$ transition. The double Z-characteristic for phase transitions (called as an 2Z-characteristic for phase transitions) (Fig. 37) illustrates the two types of the metastability of synchronized flow (with respect to the $S \rightarrow F$ transition and with respect to the $S \rightarrow J$ transition) predicted in the three-phase theory (Kerner 1998b, 2004a).

These two types of the metastability of synchronized flow are responsible for the $F \rightarrow S$, $S \rightarrow F$, $S \rightarrow J$, and $J \rightarrow S$ phase transitions between the three traffic phases in vehicular traffic (Fig. 37a). In its turn, these phase transitions and their features are associated with two qualitatively different instabilities in synchronized flow: (i) the $S \rightarrow F$ instability discussed in section “ $S \rightarrow F$ Instability at Bottleneck” and (ii) the $S \rightarrow J$ instability discussed in section “Competition Between Driver Speed Adaptation with Driver Over-Deceleration: Origin of $S \rightarrow J$ Instability.” Additionally to the illustration of the phase transitions (Fig. 37a), the 2Z-characteristic for phase transitions allows us also to illustrate the $S \rightarrow F$ instability (labeled by “ $S \rightarrow F$ instability” in Fig. 37b) and the $S \rightarrow J$ instability (labeled by “ $S \rightarrow J$ instability” in Fig. 37b). Due to the two types of the metastability of synchronized flow, both the $S \rightarrow F$ instability and $S \rightarrow J$ instability exhibit the nucleation nature.

The nucleation nature of the $S \rightarrow F$ instability is as follows. If the amplitude of a local speed increase in an initial synchronized flow is small enough, no $S \rightarrow F$ instability occurs in the



Modeling Approaches to Traffic Breakdown, Fig. 37 Phase transitions in traffic flow (a) and two instabilities in synchronized flow (b) on 2Z-characteristic for phase transitions (Kerner and Klenov 2003): (a) Qualitative illustration of F → S, S → F, S → J, and J → S phase transitions between the three traffic phases F, S, and J (F – free flow phase, S – synchronized flow phase, J – wide moving jam phase); dashed curves $v_{cr, FS}^{(B)}$ and $v_{cr, SJ}$ show, respectively, the average speed within a critical nucleus for

F → S transition and the average speed within a critical nucleus for S → J transition. (b) Qualitative illustration of S → F and S → J instabilities in synchronized flow. The 2Z-characteristic results from simulations of the Kerner-Klenov stochastic microscopic three-phase model made on a single-lane road with an on-ramp bottleneck at a variable on-ramp inflow rate q_{on} and a given flow rate upstream of the on-ramp bottleneck $q_{in} = 1850$ vehicles/h

synchronized flow. The S → F instability is realized only if the amplitude of a local speed *increase* in the synchronized flow is equal to or exceeds some critical amplitude. In this case, the S → F instability occurs leading to a growing wave of the local increase in the speed in synchronized flow (Figs. 34 and 35). The subsequent uninterrupted growth of this wave of the local speed increase in synchronized flow results in the S → F transition. The local speed increase in the synchronized flow initiating the S → F instability can be considered a nucleus for the S → F instability. The nucleation nature of the S → F instability is associated with a spatiotemporal competition between the over-acceleration effect and the speed adaptation effect discussed in section “[Competition Between Speed Adaptation with Over-Acceleration](#).” The nucleation nature of the S → F instability governs the nucleation nature of traffic breakdown (F → S transition) (see more detailed explanations in the book (Kerner 2017a)).

The nucleation nature of the S → J instability is as follows. If the amplitude of a local speed

decrease in an initial synchronized flow is small enough, no S → J instability occurs in the synchronized flow. The S → J instability is realized only if the amplitude of a local speed *decrease* in the synchronized flow is equal to or exceeds some critical amplitude. In this case, the S → J instability occurs leading to a growing wave of the local decrease in the speed in synchronized flow. The subsequent uninterrupted growth of this wave of the local speed reduction in synchronized flow results in the formation of a wide moving jam (S → J transition). The local speed decrease in the synchronized flow initiating the S → J instability can be considered a nucleus for the S → J instability. The nucleation nature of the S → J instability is associated with a spatiotemporal competition between the over-deceleration effect and the speed adaptation effect discussed in section “[Competition between Driver Speed Adaptation with Driver Over-Deceleration: Origin of S → J Instability](#).” In more details, the nucleation nature of the S → J instability will be considered in the article (Kerner 2018) of this Encyclopedia.

Evaluation of on-Ramp Metering with Three-Phase Theory

In this section, we illustrate the application of traffic flow modeling in the framework of the three-phase theory for the evaluation of on-ramp metering at on-ramp bottlenecks.

Firstly, we consider an application of the congested pattern control approach discussed in the article (Kerner 2017b) of this Encyclopedia for on-ramp metering called “automatic on-ramp control of congested patterns” (ANCONA) (Kerner 2004a, 2005, 2007a, b, c, d).

Then, we compare ANCONA on-ramp metering with a well-known ALINEA (Asservissement Linéaire d'Entrée Autoroutière) on-ramp metering of Papageorgiou et al. (1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003). In ALINEA control rules, it is assumed that there is a particular highway capacity of free flow at a bottleneck. In other words, ALINEA is associated with the classical understanding of highway capacity that criticism has been made in section “Failure of Classical Understanding of Stochastic Highway Capacity.” Because the classical understanding of highway capacity is inconsistent with the empirical nucleation nature of traffic breakdown, we will find that the ALINEA on-ramp metering is also inconsistent with the empirical nucleation nature of traffic breakdown that is the empirical fundamental of transportation science.

A comparison of ANCONA and ALINEA applications made below should illustrate the failure of intelligent transportation systems based on classical traffic flow theories as discussed in the article (Kerner 2017b) of this Encyclopedia.

ANCONA On-Ramp Metering

In accordance with the idea of the congested pattern approach, ANCONA does not restrict the on-ramp inflow rate q_{on} as long as free flow is measured at the bottleneck: There is no on-ramp control as long as free flow is measured at the bottleneck. This means that the flow rate of

vehicles denoted by $q_{on}^{(cont)}$, which can merge from the on-ramp onto the main road (Fig. 38a), is equal to

$$q_{on}^{(cont)} = q_{on}. \quad (120)$$

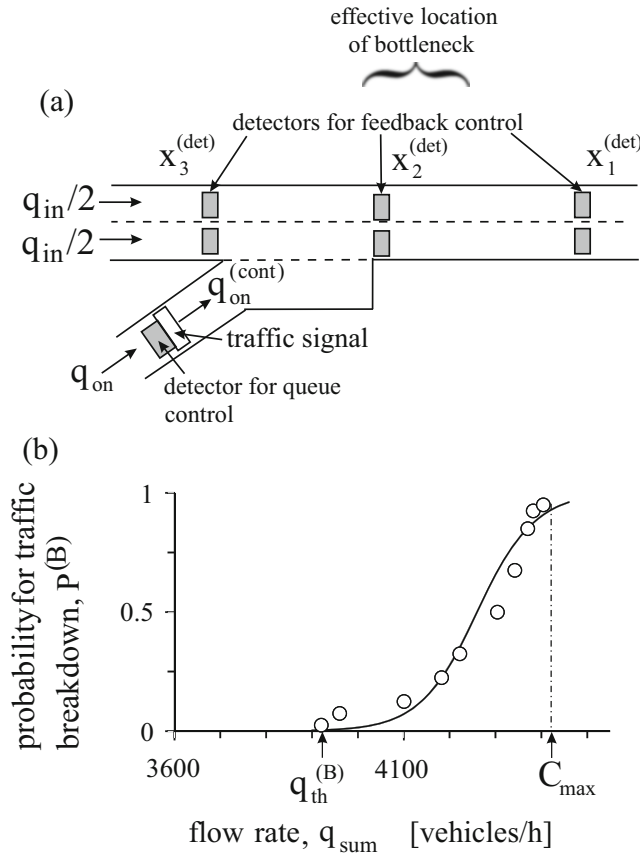
Only after synchronized flow due to random traffic breakdown is set in at the bottleneck, ANCONA begins to control this synchronized flow through the feedback decrease in the on-ramp inflow.

To measure a random traffic breakdown with the subsequent formation of synchronized flow at the bottleneck, there are feedback detectors on the main road in ANCONA. As explained in (Kerner 2004a, 2005, 2007a, b, c, d), the shortest queue at the signal in the on-ramp lane that can occur under ANCONA application is realized if the feedback detectors are located at an effective location of the bottleneck. The effective bottleneck location is a road location at which the average speed within a permanent local disturbance at the bottleneck reaches a minimum value (some of the possible locations of the effective bottleneck location are marked by a curly bracket labeled by “effective location of bottleneck” in Fig. 38a). However, for real installations it is not easy to determine the effective bottleneck location. Therefore, when the feedback detectors are located upstream of the beginning of the on-ramp merging region, a *reliable* detection of an F $\rightarrow$ S transition at the bottleneck can be expected. For this reason, in simulations, we have used the feedback detectors whose location $x_3^{(det)}$ is about 200–300 m upstream of the effective location of the bottleneck (Fig. 38a).

Due to traffic breakdown, the average speed denoted by $v^{(det)}$ measured at the feedback control detectors upstream of the bottleneck at location $x_3^{(det)}$ (Fig. 38a) decreases. If the average speed $v^{(det)}$ is equal to or drops below a chosen “congestion speed” denoted by $v_{cong, 1}$, i.e.,

$$v^{(det)} \leq v_{cong, 1}, \quad (121)$$

feedback on-ramp metering via ANCONA begins to perform. We have applied the following feedback control procedure: Through the use of traffic signal



Modeling Approaches to Traffic Breakdown, Fig. 38 Explanation of models of feedback on-ramp metering used in simulations: (a) Qualitative scheme of two-lane highway with on-ramp bottleneck with a detector in the on-ramp lane for queue control and with three different locations of feedback control detectors on the highway for feedback on-ramp metering: (i) detector location $x_1^{(det)} = 16.3$ km is related to a downstream road location far enough from the end of the merging region of on-ramp bottleneck at $x = 15.3$ km, (ii) detector location $x_2^{(det)} = 15.35$ km is related to the effective location of the bottleneck that is within the permanent speed disturbance

on the bottleneck, and (iii) detector location $x_3^{(det)} = 14.9$ km is related to the road location 100 m upstream of the beginning of the merging region of the on-ramp that is at road location $x_{on} = 15$ km. (b) Probability of traffic breakdown at the on-ramp bottleneck as function of the flow rate downstream of the bottleneck $q_{sum} = q_{in} + q_{on}$ at free flow conditions at the bottleneck and a given flow rate to the on-ramp $q_{on} = 700$ vehicles/h calculated with the Kerner-Klenov stochastic microscopic three-phase traffic flow model under condition that no on-ramp metering is applied. Maximum capacity is equal to $C_{max} = 4404$ vehicles/h

in the on-ramp line, ANCONA reduces the flow rate $q_{on}^{(cont)}$ to a smaller flow rate denoted by q_{on1} :

$$q_{on}^{(cont)} = q_{on1} < q_{on}. \quad (122)$$

Due to this decrease in the flow rate (122), ANCONA tries to achieve a return phase transition from synchronized flow to free flow ($S \rightarrow F$ transition) at the bottleneck. As a result of this $S \rightarrow F$

transition, the initial synchronized flow dissolves and the average speed at the detector $v^{(det)}$ increases above some chosen value of the speed $v_{cong, 2}$:

$$v^{(det)} > v_{cong, 2}. \quad (123)$$

When condition (123) is satisfied, through the change in parameters of traffic signal in the on-ramp lane a larger on-ramp inflow rate

$$q_{\text{on}}^{(\text{cont})} = q_{\text{on } 2}, \quad (124)$$

is allowed, where the flow rate $q_{\text{on } 2} > q_{\text{on } 1}$.

During a time interval within which a smaller on-ramp inflow rate $q_{\text{on } 1} < q_{\text{on}}$ (122) is permitted to merge onto the main road due to the signal in the on-ramp lane (Fig. 38a), a vehicle queue is built at the signal in the on-ramp lane. To dissolve the queue under condition (123), ANCONA tries to apply the on-ramp inflow rate $q_{\text{on } 2}$ (124) that is larger than the flow rate q_{on} upstream of the queue. The on-ramp flow rate $q_{\text{on } 2} > q_{\text{on}}$ remains as long as there is still a vehicle queue at the signal. The queue at the signal has been registered through the use of a detector in the on-ramp lane (labeled by “detector for queue control” in Fig. 38a). After the queue has dissolved, there is no traffic signal control in the on-ramp lane any more, i.e., ANCONA is switched off and condition (120) is satisfied.

If under condition (120) traffic breakdown occurs at the bottleneck later once more, an incipient SP begins to propagate upstream of the bottleneck. As a result, the speed at the feedback control detector decreases. Therefore, condition (121) is satisfied once more. This leads to a new decrease in $q_{\text{on}}^{(\text{cont})}$, and so on.

For simulations of applications of ANCONA, we use the Kerner-Klenov stochastic three-phase traffic flow model (Kerner and Klenov 2002, 2009b), in which the maximum speed of vehicles in free flow is a function of the space gap g between vehicles given by formula:

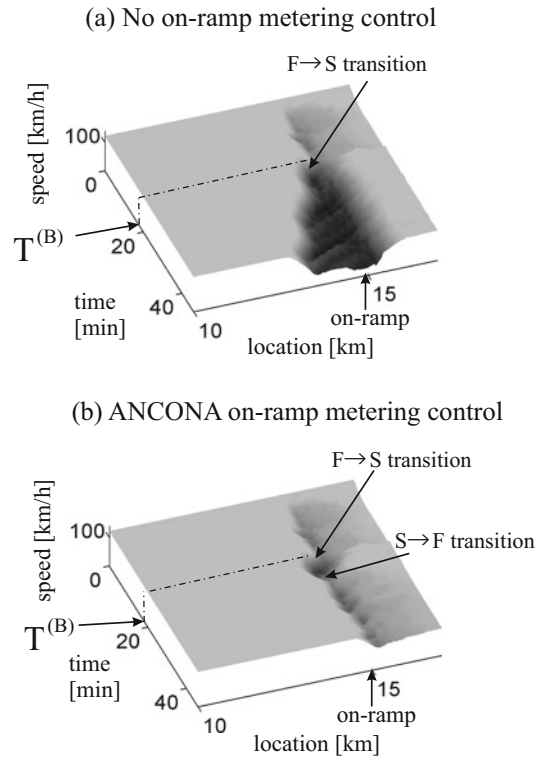
$$v_{\text{free}}(g) = \max\left(v_{\text{free}}^{(\text{min})}, v_{\text{free}}^{(\text{max})}(1 - \kappa d/(d + g))\right), \quad (125)$$

where in simulations presented below we have used the following model parameters: $v_{\text{free}}^{(\text{min})} = 90$ km/h, $v_{\text{free}}^{(\text{max})} = 150$ km/h, $d = 7.5$ m (vehicle length), and $\kappa = 1.73$.

Simulations of ANCONA application are shown in Figs. 39 and 40. When no on-ramp metering is applied, after a random time delay $T^{(B)} \approx 18$ min traffic breakdown occurs at the bottleneck (traffic breakdown is labeled by

“F $\rightarrow$ S transition” in Fig. 39a). Due to traffic breakdown, synchronized flow appears at the bottleneck. When no ANCONA application is applied, the upstream front of synchronized flow propagates continuously upstream (Fig. 39a).

Under ANCONA application (Fig. 39b), firstly, after the same random time delay $T^{(B)} \approx 18$ min traffic breakdown occurs at the bottleneck (traffic breakdown is labeled by “F $\rightarrow$ S transition” in Fig. 39b). This is because



Modeling Approaches to Traffic Breakdown, Fig. 39 Simulations of ANCONA application with the Kerner-Klenov stochastic microscopic model on a two-lane road with on-ramp bottleneck (Fig. 38a): Speed as a function of time and space without on-ramp metering (a) and under ANCONA application (b). $q_{\text{on}} = 700$ vehicles/h, $q_{\text{in}} = 3546$ vehicles/h (i.e., $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}} = 4246$ vehicles/h), probability of traffic breakdown is equal to $P^{(B)} = 0.44$. Parameters of ANCONA in (b) are $q_{\text{on } 1} = q_{\text{on}} - \Delta q_{\text{on}}^{(\text{FS})}$, $q_{\text{on } 2} = q_{\text{on}} + \Delta q_{\text{on}}^{(\text{SF})}$, $(\Delta q_{\text{on}}^{(\text{FS})}, \Delta q_{\text{on}}^{(\text{SF})}) = (250, 200)$ vehicles/h. $v_{\text{cong } 1} = v_{\text{cong } 2} = 72$ km/h. Feedback control detectors upstream of the bottleneck are at location $x_3^{(\text{det})} = 14.9$ km in Fig. 38a. The beginning of the merging region of the on-ramp bottleneck is at road location $x_{\text{on}} = 15$ km

ANCONA does not reduce the on-ramp inflow as long as free flow is realized at the bottleneck. Only after traffic breakdown has occurred, ANCONA begins to reduce the on-ramp inflow as follows. Due to traffic breakdown, synchronized flow appears upstream of the bottleneck. This leads to a decrease in the speed $v^{(\text{det})}$ at location $x_3^{(\text{det})}$ of road detectors upstream of the bottleneck (Fig. 38a). The speed value $v^{(\text{det})}$ satisfies condition (121). Therefore, feedback on-ramp metering via ANCONA begins to perform: The flow rate $q_{\text{on}}^{(\text{cont})}$ reduces to $q_{\text{on } 1}$ (122). This leads to a queue formation at the signal because the flow rate $q_{\text{on}} > q_{\text{on } 1}$ (Fig. 40a).

Due to the reducing of the on-ramp inflow through ANCONA application, the upstream front of synchronized flow does not propagate continuously upstream (Fig. 39b) as it has been in the case when no on-ramp metering has been applied (Fig. 39a). Instead, synchronized flow is now localized in a small vicinity of the bottleneck (short time interval between time instants shown

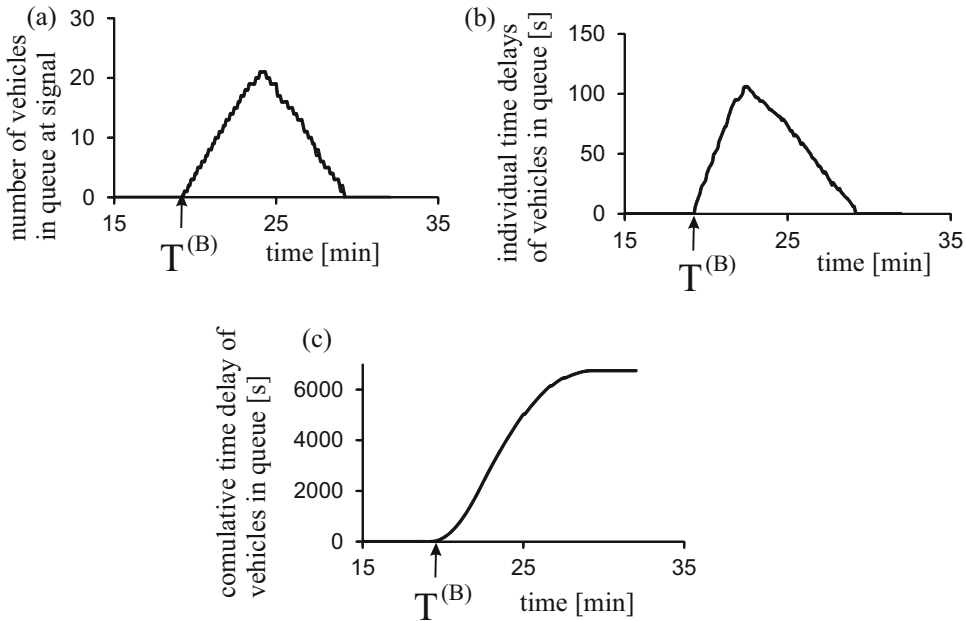
by arrows labeled by “F $\rightarrow$ S transition” and “S $\rightarrow$ F transition,” respectively, Fig. 39b).

Moreover, due to the decrease in the on-ramp inflow rate, ANCONA forces a return S $\rightarrow$ F transition at the bottleneck (labeled by “S $\rightarrow$ F transition” in Fig. 39b). As a result of this S $\rightarrow$ F transition, the initial synchronized flow dissolves and the average speed at the detector $v^{(\text{det})}$ increases above the speed $v_{\text{cong}, 2}$ (123).

When condition (123) is satisfied, through the change in parameters of traffic signal in the on-ramp lane, a larger on-ramp inflow rate $q_{\text{on}}^{(\text{cont})}$ given by Eq. 124 is allowed. For this reason, the vehicle queue is dissolved (Fig. 40a, b). Both the number of vehicles in the queue and the individual time delay of vehicles at the signal firstly increases and then decreases over time (Fig. 40a, b).

ALINEA On-Ramp Metering

The ALINEA control rule for on-ramp metering, which determines the flow rate of vehicles $q_{\text{on}}^{(\text{cont})}$



Modeling Approaches to Traffic Breakdown,
Fig. 40 Simulated characteristics of ANCONA application shown in Fig. 39b: Time-functions of number of vehicles in queue (a) and individual time delays of vehicles in the queue (b) and cumulative time delay of vehicles in

the queue (c) at the signal in the on-ramp lane. In (b), each of the points gives the individual time delay of a vehicle in [s] as a function of the time instant (time-axis) at which the vehicle comes to a stop at the end of the queue

that can merge onto the main road from the on-ramp at time t_s , $t_s = T_{av}s$, $s = 1, 2, \dots$, i.e., during the time interval $t_s \leq t < t_{s+1}$ reads as follows (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003):

$$q_{on}^{(cont)}(t_s) = q_{on}^{(cont)}(t_{s-1}) + K_o \left(o_{sum}^{(opt)} - o_{sum}^{(cont)}(t_s) \right), \quad (126)$$

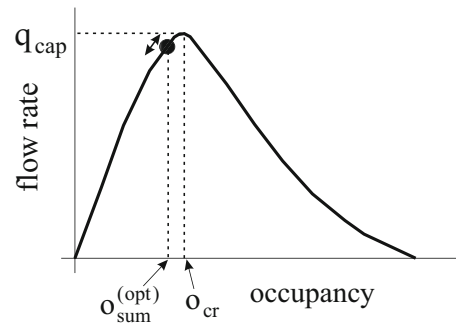
where T_{av} is the averaging time interval for data measured by feedback control detector (labeled “detectors for feedback control” in Fig. 38a); $o_{sum}^{(opt)}$ is a chosen optimal (target) occupancy; $o_{sum}^{(cont)}(t_s)$ is occupancy measured via the feedback control detectors and averaged during the time interval $t_{s-1} < t \leq t_s$; K_o is constant (in accordance with (Papageorgiou et al. 1997; Smaragdis and Papageorgiou 2003) in all simulations we use $T_{av} = 1$ min, $K_o = 70$ vehicles/h). It should be noted that in ALINEA there is a minimum admissible on-ramp inflow rate $q_{on}^{(min)}$ and a maximum onramp inflow rate $q_{on}^{(max)}$. If $q_{on} < q_{on}^{(min)}$, then ALINEA does not limit the on-ramp inflow. In accordance with (Smaragdis and Papageorgiou 2003), in simulations of ALINEA we use $q_{on}^{(min)} = 400$ vehicles/h and $q_{on}^{(max)} = 1500$ vehicles/h.

As mentioned in (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003), ALINEA control rule (126) is related to classical control theory. However, the question that will be considered here whether this application of classical control theory (126) is consistent with the empirical nucleation nature of traffic breakdown at a highway bottleneck that is the empirical fundamental of transportation science. Below we show that ALINEA control rule (126) (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) is basically inconsistent with the empirical fundamental of transportation science.

Indeed, the understanding of the sense of the optimal occupancy $o_{sum}^{(opt)}$ in ALINEA control rule (126) and the understanding how to choose the optimal occupancy $o_{sum}^{(opt)}$ are associated with the LWR traffic flow theory (Lighthill and Whitham 1955): It is assumed in ALINEA (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003) that the capacity downstream of the bottleneck q_{cap} is found from formula (20), i.e., the capacity is related to the maximum point of the fundamental diagram for downstream free flow at the bottleneck (Fig. 41). Following (Smaragdis and Papageorgiou 2003), we denote the critical occupancy associated with this capacity q_{cap} by o_{cr} (Fig. 41). In accordance with these assumptions (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003), traffic breakdown should not occur as long as downstream occupancy $o_{sum}^{(cont)}$ does not exceed critical occupancy o_{cr} (Fig. 41). Thus, in Eq. 126 the condition

$$o_{sum}^{(opt)} < o_{cr} \quad (127)$$

should be satisfied (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and



Modeling Approaches to Traffic Breakdown, Fig. 41 Explanation of theoretical basis of ALINEA (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003): Qualitative shape of a fundamental diagram, critical occupancy o_{cr} , and optimal occupancy $o_{sum}^{(opt)}$

Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003).

To prevent congestion emergence at the bottleneck at the highest possible throughput downstream of the bottleneck, in accordance with (Papageorgiou et al. 1990a, b, 1991, 1997, 2003, 2007; Papageorgiou and Kotsialos 2000, 2002; Papageorgiou and Papamichail 2008; Smaragdis and Papageorgiou 2003), in the ALINEA rule (126) the optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ should be chosen as close as possible to o_{cr} :

$$\frac{o_{\text{cr}} - o_{\text{sum}}^{(\text{opt})}}{o_{\text{sum}}^{(\text{opt})}} \ll 1. \quad (128)$$

Through the use of ALINEA, occupancy $o_{\text{sum}}^{(\text{cont})}(t_s)$ measured downstream of the bottleneck can change over time; however, this occupancy should be in a neighborhood of $o_{\text{sum}}^{(\text{opt})}$ (labeled by black point in Fig. 41) (Smaragdis and Papageorgiou 2003). One might indeed conclude that ALINEA strategy ensures the highest possible throughput downstream of the bottleneck and it maintains free flow at the bottleneck.

Criticism on ALINEA On-Ramp Metering

For simulations of applications of ALINEA, we use the Kerner-Klenov stochastic three-phase traffic flow model (Kerner and Klenov 2002, 2009b) with the same model parameters as those used in section “ANCONA On-Ramp Metering” for simulations of applications of ANCONA.

The criticism of ALINEA applications discussed below in details can be formulated as follows. Depending on the chosen optimal occupancy and the location of the feedback control detectors different applications of ALINEA (ALINEA, UP-ALINEA, ALINEA/Q, X-ALINEA/Q) (Smaragdis and Papageorgiou 2003) show one of the following two main results:

1. ALINEA fails to prevent traffic breakdown.
2. ALINEA can prevent traffic breakdown at the bottleneck. However, the prevention of the breakdown through ALINEA occurs *only* at the expense of a growing vehicle queue at traffic signal in the on-ramp lane.

Recall that at a given time-independent on-ramp flow rate q_{on} ANCONA application can lead to the dissolution of the vehicle queue at the signal (Fig. 40). At the same on-ramp flow rates q_{on} and q_{in} as those in ANCONA application (Fig. 40), we will find that at any choices of the optimal occupancy and location of feedback control detectors, at which one of the ALINEA applications can prevent traffic breakdown, the ALINEA application does cause the growing vehicle queue at the signal that cannot be dissolved.

The choice of the optimal occupancy, location of the feedback control detectors, and any other parameters of ALINEA effects only on whether result 1 or result 2 of ALINEA application is realized. The results 1 and 2 are the consequence of the inconsistency of ALINEA with the empirical nucleation nature of traffic breakdown. Indeed, the theoretical basis of ALINEA – traffic breakdown at the bottleneck does not occur as long as condition (127) is satisfied (Fig. 41) – is inconsistent with the empirical nucleation nature of traffic breakdown (section “Empirical Proof of Nucleation Nature of Traffic Breakdown (F→S Transition)”). In the remaining of section “ALINEA On-Ramp Metering,” we illustrate these critical conclusions and explain their physical meaning.

Choice of Location of Feedback Control Detectors and Optimal Occupancy

In the three-phase theory, the analog of classical capacity q_{cap} (Fig. 41) is the flow rate in free flow downstream of the bottleneck $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ that is equal to the maximum highway capacity C_{max} (Fig. 38b). For this reason, in simulations the critical occupancy o_{cr} of the classical theory (Fig. 41) is related to the occupancy in free flow at the bottleneck found from condition

$$o_{\text{cr}} = o(C_{\text{max}}). \quad (129)$$

Indeed, in the classical theory when the flow rate q_{sum} is larger than q_{cap} , traffic breakdown does occur at the bottleneck (Fig. 41). The same assumption is made in the three-phase theory:

When q_{sum} is larger than C_{max} , then traffic breakdown does occur at the bottleneck during a chosen observation time T_{ob} (in simulations, we use $T_{\text{ob}} = 30$ min) with the probability $P^{(\text{B})}$ that is equal to 1 (Fig. 38b). The basic difference between the three-phase theory and the classical theories is that in the three-phase theory, when the flow rate q_{sum} satisfies conditions (Fig. 38b)

$$q_{\text{th}}^{(\text{B})} \leq q_{\text{sum}} < C_{\text{max}}, \quad (130)$$

then traffic breakdown can occur at the bottleneck with the probability $0 < P^{(\text{B})}(q_{\text{sum}}) < 1$, whereas in the classical theories under condition (Fig. 41)

$$q_{\text{sum}} < q_{\text{cap}} \quad (131)$$

no traffic breakdown can occur.

It should be noted that in Eq. 130 the flow rate $q_{\text{sum}} = q_{\text{th}}^{(\text{B})}$ is a threshold flow rate for spontaneous traffic breakdown at the bottleneck: Although under conditions $C_{\text{min}} \leq q_{\text{sum}} < q_{\text{th}}^{(\text{B})}$ free flow is in a metastable state with respect to $F \rightarrow S$ transition at the bottleneck (see Fig. 16), the probability that traffic breakdown occurs at the bottleneck during the observation time T_{ob} is equal to zero, i.e., $P^{(\text{B})}(q_{\text{sum}}) = 0$. In this case, traffic breakdown can only be induced at the bottleneck. A more detailed consideration of the threshold flow rate for spontaneous traffic breakdown at the bottleneck can be found in Sect. 5.4 of the book (Kerner 2017a).

In earlier ALINEA applications, to measure the occupancy $o_{\text{sum}}^{(\text{cont})}(t_s)$ associated with the sum of the flow rates $q_{\text{in}} + q_{\text{on}}$ and to avoid the influence of disturbances due to vehicle merging within the on-ramp merging region, the feedback control detectors in ALINEA of Ref. (Papageorgiou et al. 1990a, b) are located at some distance downstream of the end of the on-ramp merging region (location of feedback detectors $x_1^{(\text{det})}$ in Fig. 38a).

Later UP-ALINEA has been developed (Smaragdis and Papageorgiou 2003) in which the feedback control detectors are upstream of the bottleneck (location of feedback detectors $x_3^{(\text{det})}$ in Fig. 38a). Simulations of UP-ALINEA can be made with ALINEA control rules (126) in

which in accordance with (Smaragdis and Papageorgiou 2003) we use

$$o_{\text{sum}}^{(\text{cont})}(t_s) = o_{\text{in}}(t_s) \left[1 + \frac{\tilde{q}_{\text{on}}^{(\text{cont})}(t_s)}{q_{\text{in}}(t_s)} \right], \quad (132)$$

where o_{in} is the occupancy measured by feedback control detectors at location $x_3^{(\text{det})}$, $\tilde{q}_{\text{on}}^{(\text{cont})}$ is the measured on-ramp flow entering the freeway during the time interval $t_s - 1 < t \leq t_s$, and it has been taken into account that the lane numbers on the main road is the same upstream and downstream of the bottleneck (Fig. 38a).

To demonstrate that independent of location of feedback control detectors ALINEA and its variants (ALINEA, UP-ALINEA, X-ALINEA/Q) either cannot prevent traffic breakdown at the bottleneck or do lead to a growing vehicle queue at traffic signal in the onramp lane, we consider three locations of feedback control detectors $x_1^{(\text{det})}$, $x_2^{(\text{det})}$, and $x_3^{(\text{det})}$ (Fig. 38a). We have found that the performance of ALINEA or UP-ALINEA is qualitatively the same at any other location of feedback control detectors within the range $x_3^{(\text{det})} \leq x \leq x_1^{(\text{det})}$.

At the chosen simulation parameters, the maximum highway capacity is equal to constant value $C_{\text{max}} = 4404$ vehicles/h (Fig. 38b). It should be noted that as in real free flow at a highway bottleneck, the Kerner-Klenov model used for simulations shows the existence of a permanent speed disturbance at the bottleneck. Both the spatial distribution of the average speed and the location of the minimum speed (effective location of the bottleneck, Fig. 38a) within the permanent speed disturbance depend on the flow rate q_{sum} . Due to the spatial dependence of the average speed in free flow at the bottleneck, at the same flow rate $q_{\text{sum}} = C_{\text{max}} = 4404$ vehicles/h the value of the critical occupancy in Eq. 129 depends considerably on the location of feedback control detectors (see captions to Figs. 43, 44, and 45).

It should be noted that all simulations of on-ramp metering are made here at the same flow rate $q_{\text{on}} = 700$ vehicles/h. For this reason, the flow rate q_{sum} changes through the change in the flow rate upstream of the bottleneck q_{in} (Fig. 38a).

For simulations of the ALINEA rule (126), we choose an optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ that satisfies conditions (127) and (128). To reach this goal, we consider the flow rate $q_{\text{sum}} = 4316$ vehicles/h that is 2% less than the maximum capacity. These values of the optimal occupancy are used below for simulations of the performance of ALINEA applications (see captions to Figs. 43, 44, and 45). However, the choice of the optimal occupancy that satisfies condition (127) does not affect on the criticism of ALINEA applications formulated in section “Criticism on ALINEA On-Ramp Metering.” To demonstrate this, additionally with the optimal occupancies mentioned above, we consider below some examples of ALINEA applications for other values of the optimal occupancies.

Traffic Breakdown under ALINEA Application

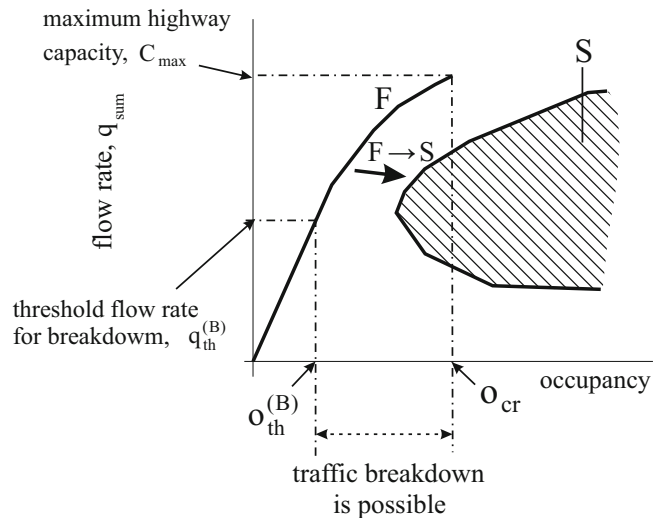
We recall that in an initially free flow downstream of the bottleneck, traffic breakdown occurs with the breakdown probability that is larger than zero when the flow rate q_{sum} is within a range (130). The flow rate (130) is related to the following range of the optimal occupancy $o_{\text{sum}}^{(\text{opt})}$ in ALINEA control rules (126):

$$o_{\text{th}}^{(\text{B})} \leq o_{\text{sum}}^{(\text{opt})} < o_{\text{cr}}, \quad (133)$$

where

Modeling Approaches to Traffic Breakdown,

Fig. 42 Explanation of failure of ALINEA to prevent traffic breakdown (Adapted from Kerner 2004a)



$$o_{\text{th}}^{(\text{B})} = o(q_{\text{th}}^{(\text{B})}) \quad (134)$$

is the value of occupancy o in free flow at location of feedback control detectors related to the flow rate q_{sum} that is equal to the threshold flow rate $q_{\text{sum}} = q_{\text{th}}^{(\text{B})}$. The range of the optimal occupancy of ALINEA application (133) is labeled in Fig. 42 as “traffic breakdown is possible.”

We can expect that far enough downstream of the effective location of the bottleneck, during traffic breakdown at the bottleneck, no noticeable change in the occupancy occurs.

Therefore, if the occupancy $o_{\text{sum}}^{(\text{cont})}$ in ALINEA control rules (126) is measured downstream of the bottleneck at road locations at which no noticeable disturbances from vehicles merging at the bottleneck are realized, as recommended in (Papageorgiou et al. 1990a, b), then with the probability $P^{(\text{B})} > 0$ traffic breakdown should occur during the time interval $T^{(\text{B})}$ at any choice of the optimal occupancy within the range (133). If the occupancy $o_{\text{sum}}^{(\text{cont})} = o_{\text{cr}}$, then the breakdown probability $P^{(\text{B})} = 1$. Thus, ALINEA control rules (126) should have no effect on the occurrence of traffic breakdown at the bottleneck.

This effect is indeed realized when the location of feedback control detectors is equal to $x_1^{(\text{det})}$. We have found that at the optimal occupancy $o_{\text{sum}}^{(\text{opt})} = 17.4\%$ in ALINEA control rules (126) that is

smaller than $o_{cr} = 17.7\%$ traffic breakdown occurs on the same way as it happens in the case when no on-ramp metering control is applied: Under ALINEA application, we have found the same occurrence of traffic breakdown with the subsequent formation of congested traffic pattern as that shown in Fig. 39a.

- If the location of feedback detectors $x_1^{(det)}$ is far enough downstream of the on-ramp merging region, then for any choice of the optimal occupancy $o_{sum}^{(opt)}$ within the occupancy range $o_{th}^{(B)} \leq o_{sum}^{(opt)} < o_{cr}$ (133), under ALINEA application traffic breakdown can occur at the bottleneck with the breakdown probability $0 < P^{(B)} < 1$ during the time interval T_{ob} .

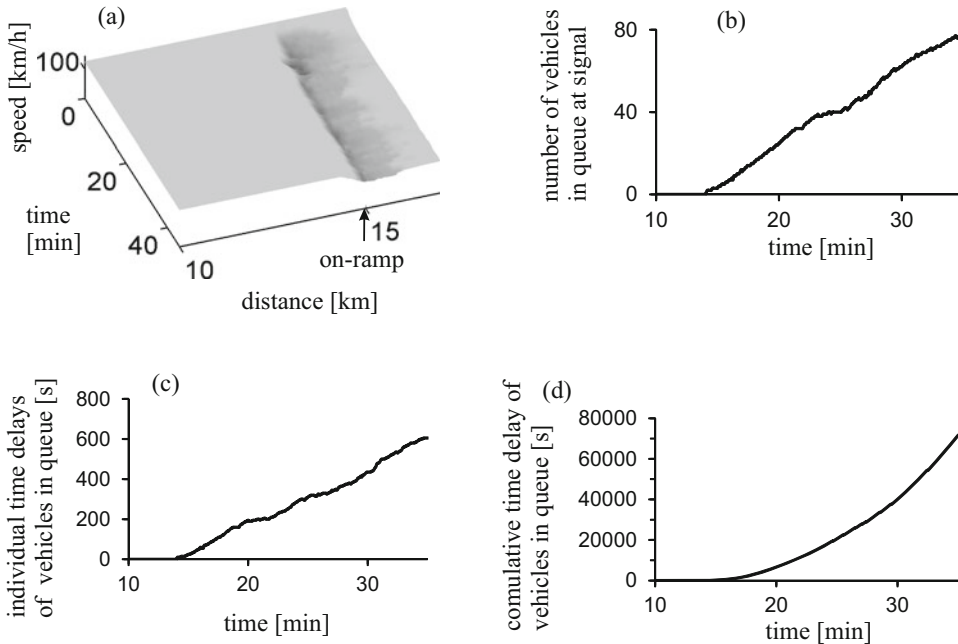
From Eq. 133 we can expect that if ALINEA is applied at the location of feedback control

detectors $x_1^{(det)}$, then the breakdown probability during the time interval T_{ob} is equal to zero only when the optimal occupancy $o_{sum}^{(opt)}$ in ALINEA control rules (126) is chosen less than the threshold occupancy $o_{th}^{(B)}$ related to the threshold flow rate for spontaneous traffic breakdown $q_{th}^{(B)}$. Indeed, we have found that when condition

$$o_{sum}^{(opt)} < o_{th}^{(B)} \quad (135)$$

is satisfied, then no traffic breakdown occurs under ALINEA application. An example is shown in Fig. 43a.

At such small optimal occupancy $o_{sum}^{(opt)} = 14.4\%$, ALINEA reduces the on-ramp inflow rate $q_{on}^{(cont)}$ through the use of traffic signal in the on-ramp lane considerably in comparison with on-ramp flow rate $q_{on} = 700$ vehicles/h. This decrease in the on-ramp inflow through ALINEA prevents the



Modeling Approaches to Traffic Breakdown, Fig. 43 ALINEA can prevent traffic breakdown at small enough chosen optimal occupancy $o_{sum}^{(opt)} = 14.4\%$ satisfying condition (135) (a) at the expense of the occurrence of a continuously growing long queue at the signal only (b–d). Detector location $x_1^{(det)}$ in Fig. 38a. (a) Speed on the main road in space and time. (b) Time-functions of number of vehicles in queue. (c) Individual time delays of vehicles in the queue. (d) Cumulative time delay of vehicles in the

queue at the signal in the on-ramp lane. In (c), each of the points gives the individual time delay of a vehicle in [s] as a function of the time instant (time-axis) at which the vehicle comes to a stop at the end of the queue. The flow rate is the same as those in Fig. 39a: $q_{on} = 700$ vehicles/h, $q_{sum} = q_{in} + q_{on} = 4246$ vehicles/h. In accordance with (Smaragdis and Papageorgiou 2003), we use $q_{on}^{(min)} = 400$ vehicles/h

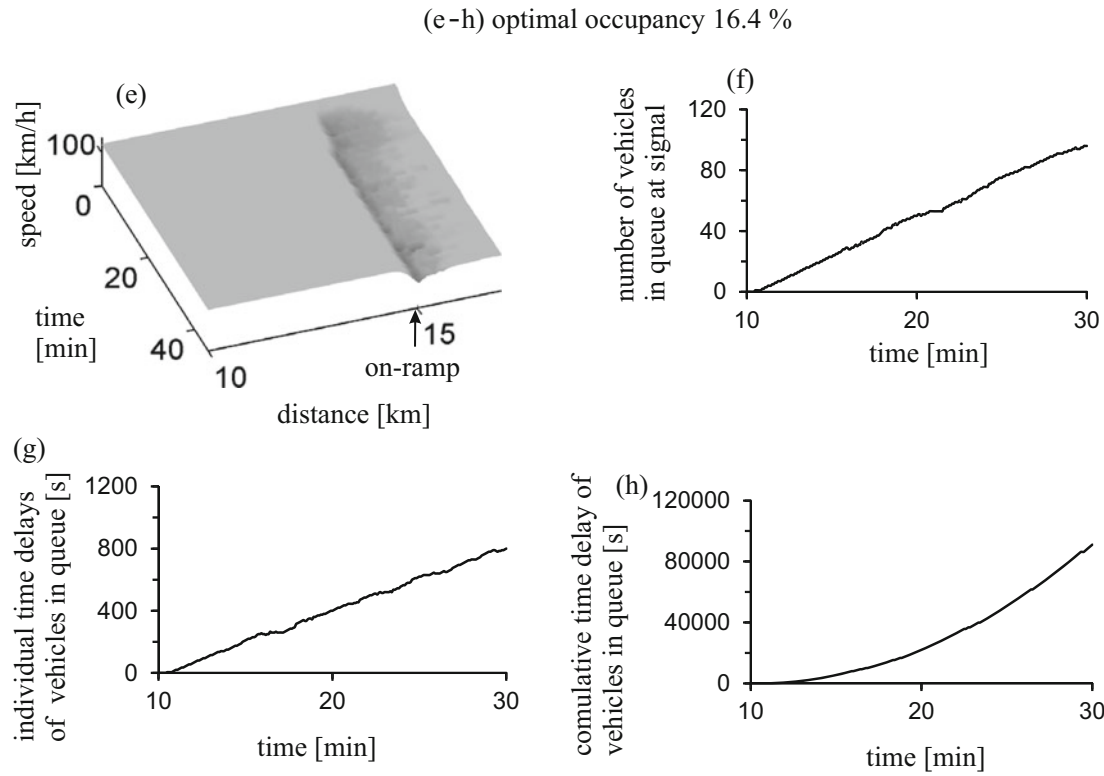
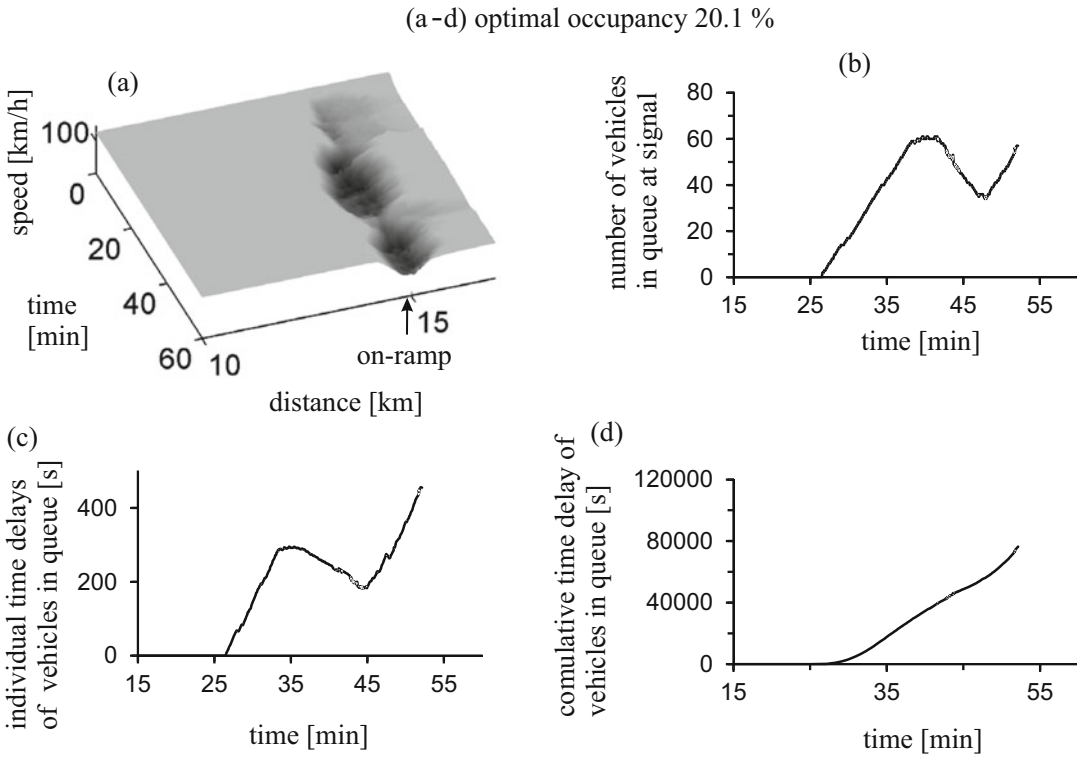
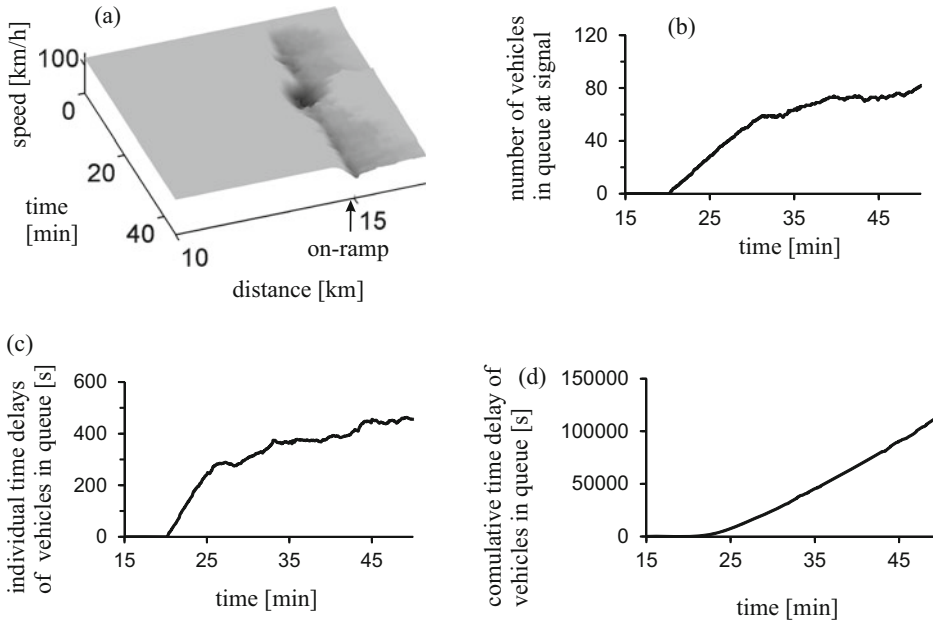


Fig. 44 (continued)

occurrence of traffic breakdown (Fig. 43a). However, this prevention of traffic breakdown is realized at the expense of the occurrence of the continuously growing long queue at the signal only (Fig. 43b–d). We see that the total delay time of vehicles at the signal becomes about 10 times larger (Fig. 43d) as that has been under ANCONA application (Fig. 40c).



Modeling Approaches to Traffic Breakdown, Fig. 45 UP-ALINEA can prevent traffic breakdown (a), however, at the expense of the occurrence of growing long queue at the signal only (b–d). Location of feedback detectors $x_3^{(det)} = 14.9$ km in Fig. 38a. (a) Speed on the main road in space and time. (b) Time-functions of number of vehicles in queue. (c) Individual time delays of vehicles in the queue. (d) Cumulative time delay of vehicles in the queue at the signal in the on-ramp lane. In (c),

Reduction of Congestion or Prevention of Traffic Breakdown under ALINEA Application

Both empirical results and simulations with the Kerner-Klenov model show that during traffic breakdown the speed decreases and the occupancy increases. This occurs firstly at the effective location of the bottleneck and later upstream of the bottleneck. For this reason, we

each of the points gives the individual time delay of a vehicle in [s] as a function of the time instant (time-axis) at which the vehicle comes to a stop at the end of the queue. The optimal occupancy is equal to $o_{sum}^{(opt)} = 17.8\%$ that is related to the flow rate q_{sum} which is 2% less than the maximum capacity at which $o_{cr} = 19\%$ for the given detector location $x_3^{(det)} = 14.9$ km. The flow rate are the same as those in Fig. 39a: $q_{on} = 700$ vehicles/h, $q_{sum} = q_{in} + q_{on} = 4246$ vehicles/h

Modeling Approaches to Traffic Breakdown, Fig. 44 ALINEA can either reduce traffic congestion (a–d) or prevent traffic breakdown (e–h) at location of feedback control detectors $x_2^{(det)} = 15.35$ km. (a, e) Speed on the main road in space and time. (b, f) Time-functions of number of vehicles in queue. (c, g) Individual time delays of vehicles in the queue. (d, h) Cumulative time delay of vehicles in the queue at the signal in the on-ramp lane. In (c, g), each of the points gives the individual time delay of a vehicle in [s] as a function of the time instant (time-axis) at which the vehicle comes to a stop at the end of the queue.

In (a–d), the optimal occupancy is equal to $o_{sum}^{(opt)} = 20.1\%$ that is related to the flow rate q_{sum} which is 2% less than the maximum capacity at which $o_{cr} = 20.7\%$ for the given detector location $x_2^{(det)} = 15.35$ km. In (e–h) the optimal occupancy is equal to $o_{sum}^{(opt)} = 16.4\%$. Both the reduction of congestion (a) and the prevention of traffic breakdown (e) are possible only at the expense of the occurrence of growing long queue at the signal (b–d, f–h). The flow rate are the same as those in Fig. 39a: $q_{on} = 700$ vehicles/h, $q_{sum} = q_{in} + q_{on} = 4246$ vehicles/h

can expect that when the location of feedback control detectors is chosen downstream of the bottleneck, however, close enough to the effective location of the bottleneck like location $x_2^{(\text{det})}$, then ALINEA can either prevent traffic breakdown or at least reduce the effect of congestion at the bottleneck.

These assumptions are indeed realized as shown in Fig. 44. When the optimal occupancy $o_{\text{sum}}^{(\text{opt})} = 20.1\%$ is chosen only slightly smaller than the critical occupancy $o_{\text{cr}} = 20.7\%$ for the detector location $x_2^{(\text{det})}$, then due to the reduction of the on-ramp inflow during traffic breakdown through the use of ALINEA, traffic congestion is localized in a small vicinity of the bottleneck (Fig. 44a). This means that ALINEA prevents upstream propagation of congestion.

If the optimal occupancy is reduced to the value $o_{\text{sum}}^{(\text{opt})} = 16.4\%$, then ALINEA is able to prevent traffic breakdown: No congestion occurs at the bottleneck (Fig. 44e).

However, both the reduction of congestion (Fig. 44a) and the prevention of traffic breakdown (Fig. 44e) are possible with the use of ALINEA only at the expense of the occurrence of growing long queue at the signal (Fig. 44b–d, f–h). The case of the reduction of congestion leads to a slower vehicle queue growth at the signal (Fig. 44b–d) in comparison with the case of the breakdown prevention (Fig. 44f–h). However, in both the cases, the total vehicle time delay in the queue grows unlimited over time. In contrast with the ALINEA applications, ANCONA prevents traffic breakdown with a limited queue that dissolves fully over time (Fig. 40).

Qualitatively the same result is found when the location of feedback control detectors $x_3^{(\text{det})}$ is upstream of the bottleneck. The prevention of traffic breakdown with the use of UP-ALINEA (Fig. 45) is possible only at the expense of the occurrence of a unlimited growing long queue at the signal (Fig. 45b–d). In contrast with UP-ALINEA applications, the use of ANCONA at the same detector location prevents traffic breakdown with a limited queue that dissolves fully over time (Fig. 40).

Conclusions

1. Empirical traffic breakdown at a freeway bottleneck, i.e., the onset of congestion in an initial free flow at the bottleneck, is associated with a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition). An $F \rightarrow S$ transition at the bottleneck can be spontaneous or induced.
2. Empirical features of nuclei for traffic breakdown at highway bottlenecks confirm the basic assumption of the three-phase theory about the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway bottlenecks.
3. Classical traffic flow models used by traffic researches can be classified into two main classes referred to the classical LWR-theory and to the classical GM-model. None of the classical traffic flow models (see, e.g., traffic flow models reviewed in (Chowdhury et al. 2000; Cremer 1979; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Saifuzzaman and Zheng 2014; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999) can explain and predict the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a highway bottleneck. We call the classical traffic flow models as “two-phase models,” to distinguish them from three-phase traffic flow models developed in the framework of the three-phase theory. These two-phase traffic flow models are basic traffic flow models for simulations of highway control and management strategies as well as other intelligent transportation systems (ITS). We have to conclude that the related simulations of the control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated management strategy. This explains the failure of applications of the classical traffic flow theories and models for the development and simulations of reliable ITS, including simulations of the effect of autonomous driving vehicles on traffic flow.

4. In contrast with the two-phase traffic flow models and classical traffic flow theories (see, e.g., traffic flow models and theories reviewed in (Chowdhury et al. 2000; Cremer 1979; Cowan 1976; Daganzo 1997; Gartner et al. 1997; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005; May 1990; Nagatani 2002; Nagel et al. 2003; Saifuzzaman and Zheng 2014; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974; Wolf 1999), three-phase traffic flow models developed in the framework of the three-phase theory can show and explain the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition). Thus, three-phase traffic flow models should be used for reliable simulations of control and management traffic strategies and other ITS before they are introduced to the market.

Future Directions

There are at least the following challenging areas for traffic modeling in which the three-phase theory should be used in the future:

- (i) The development of new traffic flow models in the framework of the three-phase theory should be intensified. This is because only three-phase traffic flow models that incorporate the metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck can be used for reliable analysis of the performance of ITS in traffic flow.
- (ii) Congested pattern control approach based on the three-phase theory should be further developed and implemented to the market.
- (iii) The three-phase theory should be also a theoretical basis for the future development of effective methods of traffic control in networks in the case of an extremely large total network inflow rate at which congestion propagates through the whole network. Indeed, in many real traffic networks, there are not enough alternative routes to avoid traffic congestion at large enough traffic demand. Thus, at large enough traffic demand, methods of traffic control and dynamic network optimization could not prevent traffic breakdown. Nevertheless, the application of ITS can change characteristics of traffic congestion with the objective to increase traffic safety and comfort while moving in congested traffic.
- (iv) The development of automated driving based on the three-phase theory whose dynamic behavior is automatically adaptable to dynamic behavior of manual drivers should be performed.
- (v) There are almost no analytical results within the context of the three-phase theory (the exception is a probabilistic theory of traffic breakdown at a bottleneck (see Kerner and Klenov 2009a, 2018)). These and many other unsolved problems are an interesting and important fields of the future empirical and theoretical investigations in traffic and transportation research (for more details, see chapter “► [Modelling of Pedestrian and Evacuation Dynamics](#)” of the book (Kerner 2017a)).
- (vi) Studies of microscopic (single vehicle) empirical features of traffic breakdown at network bottlenecks and resulting congested patterns should be intensified. In particular, the empirical studies should give the answer to questions about features of congested traffic that can effectively be influenced through the use of ITS-applications for the spatial limitation of congestion growth and/or for the dissolution of traffic congestion.
- (vii) There are FOTO and ASDA models for tracking of spatiotemporal congested patterns based on the three-phase theory, which are already successfully used in on-line installations (see results and references in (Hausken and Rehborn 2015; Kerner 2004a; Rehborn and Klenov 2009; Rehborn et al. 2011a, b, 2017; Rehborn and Koller 2014; Rehborn and Palmer 2008; Rempe et al. 2016, 2017)). We can expect that many other applications for control and management of traffic networks in the framework of the three-phase theory will be developed in the future.

Acknowledgments I would like to thank Sergey Klenov for help and useful suggestions. We thank our partners for their support in the project “MEC-View – Object detection for automated driving based on Mobile Edge Computing,” funded by the German Federal Ministry of Economic Affairs and Energy.

Bibliography

- Ahn S, Cassidy MJ (2007) Freeway traffic oscillations and vehicle lane-change maneuvers. In: Allsop RE, Bell MGH, Hydecker BG (eds) *Transportation and traffic theory 2007*. Elsevier, Amsterdam, pp 691–710
- Aw A, Rascle M (2000) Resurrection of “second order” models of traffic flow. *SIAM J Appl Math* 60:916–938
- Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1994) Structure stability of congestion in traffic dynamics. *Jpn J Appl Math* 11:203–223
- Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995a) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51:1035–1042
- Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995b) Phenomenological study of dynamical model of traffic flow. *J Phys I France* 5:1389–1399
- Banks JH (1989) Freeway speed-flow-concentration relationships: more evidence and interpretations (with discussion and closure). *Transp Res Rec* 1225:53–60
- Banks JH (1990) Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering? *Transp Res Rec* 1287:20–28
- Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793–800
- Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. *Math Models Methods Appl Sci* 12:1801–1843
- Berg P, Woods A (2001) On-ramp simulations and solitary waves of a car-following model. *Phys Rev E* 64:035602(R)
- Bovy PHL (ed) (1998) *Motorway analysis: new methodologies and recent empirical findings*. Delft University Press, Delft
- Brilon W, Zurlinden H (2004) Kapazität von Straßen als Zufallsgröße. *Straßenverkehrstechnik* 4:164
- Brilon W, Geistefeld J, Regler M (2005a) Reliability of freeway traffic flow: a stochastic concept of capacity. In: Mahmassani HS (ed) *Transportation and traffic theory. Proceedings of the 16th international symposium on transportation and traffic theory*. Elsevier, Amsterdam, pp 125–144
- Brilon W, Regler M, Geistefeld J (2005b) Zufallscharakter der Kapazität von Autobahnen und praktische Konsequenzen – Teil 1. *Straßenverkehrstechnik* 3:136
- Brockfeld E, Kühne RD, Skabardonis A, Wagner P (2003) Toward benchmarking of microscopic traffic flow models. *Trans Res Rec* 1852:124–129
- Brockfeld E, Kühne RD, Wagner P (2005) Calibration and validation of simulation models. In: *Proceeding of the transportation research board 84th annual meeting*, TRB paper no. 05-2152. TRB, Washington, DC
- Ceder A (ed) (1999) *Transportation and traffic theory*. In: *Proceedings of the 14th international symposium on transportation and traffic theory*. Elsevier Science Ltd, Oxford
- Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6:165–184
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
- Cowan RJ (1976) Useful headway models. *Trans Res* 9:371–375
- Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
- Daganzo CF (1994) The cell-transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B* 28:269–287
- Daganzo CF (1995) The cell transmission model, part II: network traffic. *Transp Res B* 29:79–93
- Daganzo CF (1997) *Fundamentals of transportation and traffic operations*. Elsevier Science, New York
- Daganzo CF, Cassidy MJ, Bertini RL (1999) Possible explanations of phase transitions in highway traffic. *Transp Res A* 33:365–379
- Davis LC (2004a) Multilane simulations of traffic phases. *Phys Rev E* 69:016108
- Davis LC (2004b) Effect of adaptive cruise control systems on traffic flow. *Phys Rev E* 69:066110
- Davis LC (2006a) Controlling traffic flow near the transition to the synchronous flow phase. *Physica A* 368:541–550
- Davis LC (2006b) Effect of cooperative merging on the synchronous flow phase of traffic. *Physica A* 361:606–618
- Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. *Physica A* 379:274–290
- Davis LC (2008) Driver choice compared to controlled diversion for a freeway double on-ramp in the framework of three-phase traffic theory. *Physica A* 387:6395–6410
- Davis LC (2014) Nonlinear dynamics of autonomous vehicles with limits on acceleration. *Physica A* 405:128–139
- Davis LC (2016) Improving traffic flow at a 2-to-1 lane reduction with wirelessly connected, adaptive cruise control vehicles. *Physica A* 451:320–332
- Edie LC (1961) Car-following and steady state theory for non-congested traffic. *Oper Res* 9:66–77
- Edie LC, Foote RS (1958) Traffic flow in tunnels. *Highway Res Board Proc Ann Meet* 37:334–344
- Edie LC, Foote RS (1960) Effect of shock waves on tunnel traffic flow. In: *Highway research board proceedings*, vol 39. HRB, National Research Council, Washington, DC, pp 492–505
- Edie LC, Herman R, Lam TN (1980) Observed multilane speed distribution and the kinetic theory of vehicular traffic. *Transp Sci* 14:55–76

- Elefteriadou L (2014) An introduction to traffic flow theory. Springer optimization and its applications, vol 84. Springer, Berlin
- Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Transp Res Rec* 1484:80–89
- Elefteriadou L, Kondyli A, Brilon W, Hall FL, Persaud B, Washburn S (2014) Enhancing ramp metering algorithms with the use of probability of breakdown models. *J Transp Eng* 140:04014003
- Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) *Traffic and granular flow* '01. Springer, Heidelberg
- Gao K, Jiang R, Hu S-X, Wang B-H, Wu Q-S (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. *Phys Rev E* 76:026105
- Gartner NH, Messer CJ, Rathi A (eds) (1997) Special report 165: revised monograph on traffic flow theory. Transportation Research Board, Washington, DC
- Gazis DC (2002) *Traffic theory*. Springer, Berlin
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7: 499–505
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
- Gipps PG (1981) Behavioral car-following model for computer simulation. *Trans Res B* 15:105–111
- Green Shields BD, Bibbins JR, Channing WS, Miller HH (1935) A study of traffic capacity. *Highw Res Board Proc* 14:448–477
- Haight FA (1963) *Mathematical theories of traffic flow*. Academic, New York
- Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Trans Res Rec* 1320:91–98
- Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transp Res Rec* 1365:12–18
- Hausken K, Rehborn H (2015) Game-theoretic context and interpretation of Kerner's three-phase traffic theory. In: Hausken K, Zhuang J (eds) *Game theoretic analysis of congestion, safety and security*. Springer series in reliability engineering. Springer, Berlin, pp 113–141
- He S, Guan W, Song L (2010) Explaining traffic patterns at on-ramp vicinity by a driver perception model in the framework of three-phase traffic theory. *Physica A* 389:825–836
- Hegyi A, Bellemans T, De Schutter B (2017) Freeway traffic management and control. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. Springer, Berlin
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Helbing D, Hennecke A, Treiber M (1999) Phase diagram of traffic states in the presence of inhomogeneities. *Phys Rev Lett* 82:4360–4363
- Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) *Traffic and granular flow* '99. Springer, Heidelberg
- Helbing D, Treiber M, Kesting A, Schönhof M (2009) Theoretical vs. empirical classification and prediction of congested traffic states. *Eur Phys J B* 69:583–598
- Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7:86–106
- Herrmann M, Kerner BS (1998) Local cluster effect in different traffic flow models. *Physica A* 255:163–188
- Highway Capacity Manual (2000) National Research Council, Transportation Research Board, Washington, DC
- Highway Capacity Manual (2010) National Research Council, Transportation Research Board, Washington, DC
- Hoogendoorn SP, Luding S, PHL B, Schreckenberg M, Wolf DE (eds) (2005) *Traffic and granular flow* '03. Springer, Heidelberg
- Hu X-J, Wang W, Yang H (2012) Mixed traffic flow model considering illegal lanechanging behavior: simulations in the framework of Kerner's three-phase theory. *Physica A* 391:5102–5111
- Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. *J Phys A Math Gen* 37:8197–8213
- Jiang R, Wu QS (2005) Toward an improvement over Kerner-Klenov-Wolf three-phase cellular automaton model. *Phys Rev E* 72:067103
- Jiang R, Wu QS (2007) Dangerous situations in a synchronized flow model. *Physica A* 377:633–640
- Jiang R, Hu M-B, Wang R, Wu Q-S (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner-Klenov speed adaptation model. *Phys Lett A* 365:6–9
- Jiang R, Hu MB, Zhang HM, Gao ZY, Jia B, Wu QS, Yang M (2014) Traffic experiment reveals the nature of car-following. *PLoS One* 9:e94351
- Jiang R, Hu M-B, Zhang HM, Gao ZY, Jia B, Wu QS (2015) On some experimental features of car-following behavior and how to model them. *Transp Res B* 80:338–354
- Jiang R, Jin C-J, Zhang HM, Huang Y-X, Tian J-F, Wang W, Hu M-B, Wang H, Jia B (2017) Experimental and empirical investigations of traffic flow instability. *Transp Res Proc* 23:157–173
- Kerner BS (1998a) Theory of congested traffic flow. In: Rysgaard R (ed) *Proceedings of the 3rd symposium on highway capacity and level of service*, vol 2, road directorate. Ministry of Transport, pp 621–642
- Kerner BS (1998b) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
- Kerner BS (1998c) Traffic flow: experiment and theory. In: Schreckenberg M, Wolf DE (eds) *Traffic and granular flow* '97. Springer, Singapore, pp 239–267
- Kerner BS (1999a) Congested traffic flow: observations and theory. *Trans Res Rec* 1678:160–167

- Kerner BS (1999b) The physics of traffic. *Phys World* 12:25–30
- Kerner BS (1999c) Theory of congested traffic flow: self-organization without bottlenecks. In: Ceder A (ed) *Transportation and traffic theory*. Elsevier Science, London, pp 147–171
- Kerner BS (2000) Experimental features of the emergence of moving jams in free traffic flow. *J Phys A* 33: L221–L228
- Kerner BS (2001) Complexity of synchronized flow and related problems for basic assumptions of traffic flow theories. *Netw Spat Econ* 1:35–76
- Kerner BS (2002a) Synchronized flow as a new traffic phase and related problems for traffic flow modelling. *Math Comput Model* 35:481–508
- Kerner BS (2002b) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
- Kerner BS (2004a) *The physics of traffic*. Springer, Berlin/New York
- Kerner BS (2004b) Verfahren zur Ansteuerung eines in einem Fahrzeug befindlichen verkehrsadaptiven Assistenzsystems, German patent publication DE 10308256A1. <https://google.com/patents/DE10308256A1>; Patent WO 2004076223A1 (2004) <https://google.com/patents/WO2004076223A1>; EU Patent EP 1597106B1 (2006); German patent DE 502004001669D1 (2006)
- Kerner BS (2005) Control of spatiotemporal congested traffic patterns at highway bottlenecks. *Physica A* 355:565–601
- Kerner BS (2007a) Control of spatiotemporal congested patterns at highway bottlenecks. *IEEE Trans ITS* 8:308–320
- Kerner BS (2007b) On-ramp metering based on three-phase traffic theory I. *Traffic Eng Control* 48:28–35
- Kerner BS (2007c) On-ramp metering based on three-phase theory – part III. *Traffic Eng Control* 48:68–75
- Kerner BS (2007d) Three-phase traffic theory and its applications for freeway traffic control. In: Inweldi PO (ed) *Transportation research trends*. Nova Science Publishers, New York, pp 1–97
- Kerner BS (2007e) Method for actuating a traffic-adaptive assistance system which is located in a vehicle, USA patent US 20070150167A1. <https://google.com/patents/US20070150167A1>; USA patent US 7451039B2 (2008)
- Kerner BS (2007f) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assistenzsystem, German patent publication DE 102007008253A1. <https://register.dpma.de/DPMAREgister/pat/PatSchrifteneinsicht?docId=DE102007008253A1>; German patent publication DE 102007008257A1. <https://register.dpma.de/DPMAREgister/pat/PatSchrifteneinsicht?docId=DE102007008257A1>; German patent publication DE 102007008254A1
- Kerner BS (2009a) Traffic congestion, modelling approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9302–9355
- Kerner BS (2009b) Traffic congestion, spatiotemporal features of. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9355–9411
- Kerner BS (2009c) Introduction to modern traffic flow theory and control. Springer, Berlin/New York
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Physica A* 392:5261–5282
- Kerner BS (2015a) Microscopic theory of traffic-flow instability governing traffic breakdown at highway bottlenecks: growing wave of increase in speed in synchronized flow. *Phys Rev E* 92:062827
- Kerner BS (2015b) Failure of classical traffic flow theories: a critical review. *Elektrotech Inf* 132:417–433
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Physica A* 450:700–747
- Kerner BS (2017a) *Breakdown in traffic networks: fundamentals of transportation science*. Springer, Berlin/New York
- Kerner BS (2017b) Traffic networks, breakdown in. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. Springer, Berlin. <https://doi.org/10.1007/978.3.642.27737.5701.1>
- Kerner BS (2017c) Physics of autonomous driving based on three-phase traffic theory. Springer, Berlin. arXiv:1710.10852v3. <http://arxiv.org/abs/1710.10852>
- Kerner BS (2018) Traffic congestion, spatiotemporal features of. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. Springer, Berlin
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35:L31–L43
- Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68:036130
- Kerner BS, Klenov SL (2005) Probabilistic breakdown phenomenon at on-ramps bottlenecks in three-phase traffic theory. *Cond-mat/0502281*, e-print in <http://arxiv.org/abs/cond-mat/0502281>
- Kerner BS, Klenov SL (2006a) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: congestion nucleation in spatially non-homogeneous traffic. *Physica A* 364:473–492
- Kerner BS, Klenov SL (2006b) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory. *Transp Res Rec* 1965:70–78
- Kerner BS, Klenov SL (2006c) Deterministic microscopic three-phase traffic flow models. *J Phys A Math Gen* 39:1775–1809
- Kerner BS, Klenov SL (2009a) Traffic breakdown, probabilistic theory of. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9282–9302
- Kerner BS, Klenov SL (2009b) Phase transitions in traffic flow on multilane roads. *Phys Rev E* 80:056101
- Kerner BS, Klenov SL (2018) Traffic breakdown, mathematical probabilistic approaches to. In: Meyers RA

- (ed) Encyclopedia of complexity and system science. Springer Science+Business Media LLC
- Kerner BS, Konhäuser P (1993) Cluster effect in initially homogeneous traffic flow. *Phys Rev E* 48: R2335–R2338
- Kerner BS, Konhäuser P (1994) Structure and parameters of clusters in traffic flow. *Phys Rev E* 50:54–83
- Kerner BS, Rehborn H (1996a) Experimental features and characteristics of traffic jams. *Phys Rev E* 53: R1297–R1300
- Kerner BS, Rehborn H (1996b) Experimental properties of complexity in traffic flow. *Phys Rev E* 53:R4275–R4278
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. *Phys Rev Lett* 79:4030–4033
- Kerner BS, Konhäuser P, Schilke M (1995) Deterministic spontaneous appearance of traffic jams in slightly inhomogeneous traffic flow. *Phys Rev E* 51:6243–6246
- Kerner BS, Konhäuser P, Schilke M (1996) “Dipole-layer” effect in dense traffic flow. *Phys Lett A* 215:45–56
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35:9971–10013
- Kerner BS, Klenov SL, Hiller A (2006a) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. *J Phys A Math Gen* 39:2001–2020
- Kerner BS, Klenov SL, Hiller A, Rehborn H (2006b) Microscopic features of moving traffic jams. *Phys Rev E* 73:046107
- Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. *Non Dyn* 49:525–553
- Kerner BS, Klenov SL, Schreckenberg M (2011) Simple cellular automaton model for traffic breakdown, highway capacity, and synchronized flow. *Phys Rev E* 84:046110
- Kerner BS, Klenov SL, Hermanns G, Schreckenberg M (2013a) Effect of driver overacceleration on traffic breakdown in three-phase cellular automaton traffic flow models. *Physica A* 392:4083–4105
- Kerner BS, Rehborn H, Schäfer R-P, Klenov SL, Palmer J, Lorkowski S, Witte N (2013b) Traffic dynamics in empirical probe vehicle data studied with three-phase theory: spatiotemporal reconstruction of traffic phases and generation of jam warning messages. *Physica A* 392:221–251
- Kerner BS, Klenov SL, Schreckenberg M (2014) Probabilistic physical characteristics of phase transitions at highway bottlenecks: incommensurability of three-phase and two-phase traffic-flow theories. *Phys Rev E* 89:052807
- Kerner BS, Koller M, Klenov SL, Rehborn H, Leibl M (2015) The physics of empirical nuclei for spontaneous traffic breakdown in free flow at highway bottlenecks. *Physica A* 438:365–397
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2000) Towards a realistic microscopic description of highway traffic. *J Phys A Math Gen* 33:L477–L485
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: microscopic description of traffic phases. *Phys Rev E* 65:056133
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. *Phys Rev E* 70:016115
- Kokubo S, Tanimoto J, Hagishima A (2011) A new cellular automata model including a decelerating damping effect to reproduce Kerner’s three-phase theory. *Physica A* 390:561–568
- Kometani E, Sasaki T (1958) On the stability of traffic flow. *J Oper Res Soc Jpn* 2:11–26
- Kometani E, Sasaki T (1959) A safety index for traffic with linear spacing. *Oper Res* 7:704–720
- Koshi M (2003) An interpretation of a traffic engineer on vehicular traffic flow. In: Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) *Traffic and granular flow’01*. Springer, Heidelberg, pp 199–210
- Koshi M, Iwasaki M, Ohkura I (1983) Some findings and an overview on vehicular flow characteristics. In: Hurdle VF (ed) *Proceedings of 8th international symposium on transportation and traffic theory*. University of Toronto Press, Toronto, p 403
- Krauß S, Wagner P, Gawron C (1997) Metastable states in a microscopic model of traffic flow. *Phys Rev E* 55:5597–5602
- Kuhn TS (2012) *The structure of scientific revolutions*, 4th edn. The University of Chicago Press, Chicago
- Kühne R (1991) Traffic patterns in unstable traffic flow on freeway. In: Brannolte U (ed) *Highway capacity and level of service*. A.A. Balkema, Rotterdam, pp 211–223
- Laval JA (2007) Linking synchronized flow and kinematic waves. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Traffic and granular flow’05*. Proceedings of the international workshop on traffic and granular flow. Springer, Berlin, pp 521–526
- Lee HY, Lee H-W, Kim D (1999) Dynamic states of a continuum traffic equation with on-ramp. *Phys Rev E* 59:5101–5111
- Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical restriction versus human overreaction triggering congested traffic states. *Phys Rev Lett* 92:238702
- Lesort J-B (ed) (1996) *Transportation and traffic theory*. In: Proceedings of the 13th international symposium on transportation and traffic theory. Elsevier Science Ltd, Oxford
- Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
- Li XG, Gao ZY, Li KP, Zhao XM (2007) Relationship between microscopic dynamics in traffic flow and complexity in networks. *Phys Rev E* 76:016110

- Lighthill MJ, Whitham GB (1955) On kinematic waves. I Flow movement in long rivers. II A theory of traffic flow on long crowded roads. *Proc Roy Soc A* 229:281–345
- Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. *Trans Res Cir E-C018*, pp 84–95
- Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419:1–64
- Mahmassani HS (ed) (2005) Transportation and traffic theory. In: Proceedings of the 16th international symposium on transportation and traffic theory. Elsevier, Amsterdam
- Mahnke R, Kaupuz's J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
- Mahnke R, Kaupuz's J, Lubashevsky I (2009) Physics of stochastic processes: how randomness acts in time. Wiley-VCH, Weinheim
- May AD (1990) Traffic flow fundamentals. Prentice-Hall, Englewood Cliffs
- Molzahn S-E, Kerner BS, Rehborn H, Klenov SL, Koller M (2017) Analysis of speed disturbances in empirical single vehicle probe data before traffic breakdown. *IET Intell Transp Syst* 11:604–612. <https://doi.org/10.1049/iet-its.2016.0315>
- Nagatani T (1998) Modified KdV equation for jamming transition in the continuum models of traffic. *Physica A* 261:599–607
- Nagatani T (1999) Jamming transition in a two-dimensional traffic flow model. *Phys Rev E* 59:4857–4864
- Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
- Nagatani T, Nakanishi K (1998) Delay effect on phase transitions in traffic dynamics. *Phys Rev E* 57:6415–6421
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys (France)* I 2:2221–2229
- Nagel K, Wolf DE, Wagner P, Simon P (1998) Two-lane traffic rules for cellular automata: a systematic approach. *Phys Rev E* 58:1425–1437
- Nagel K, Wagner P, Woesler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–716
- Newell GF (1961) Nonlinear effects in the dynamics of car following. *Oper Res* 9:209–229
- Newell GF (1982) Applications of queuing theory. Chapman Hall, London
- Newell GF (2002) A simplified car-following theory: a lower order model. *Transp Res B* 36:195–205
- Papageorgiou M (1983) Application of automatic control concepts in traffic flow modeling and control. Springer, Berlin/New York
- Papageorgiou M, Kotsialos A (2000) Freeway ramp metering: an overview. In: Proceedings of the 3rd annual IEEE conference on intelligent transportation systems (ITSC 2000), Dearborn, pp 228–239
- Papageorgiou M, Kotsialos A (2002) Freeway ramp metering: an overview. *IEEE Trans Intell Transp Syst* 3(4):271–280
- Papageorgiou M, Papamichail I (2008) Overview of traffic signal operation policies for ramp metering. *Transp Res Rec* 2047:28–36
- Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990a) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: Part I: Modelling. *Transp Res Part A* 24A (5):345–359
- Papageorgiou M, Blosseville J-M, Hadj-Salem H (1990b) Modelling and real-time control of traffic flow on the southern part of Boulevard Périphérique in Paris: Part II: Coordinated on-ramp metering. *Transp Res Part A* 24A (5):361–370
- Papageorgiou M, Hadj-Salem H, Blosseville J-M (1991) ALINEA: a local feedback control law for on-ramp metering. *Transp Res Rec* 1320:58–64
- Papageorgiou M, Hadj-Salem H, Middelham F (1997) ALINEA local ramp metering summary of field results. *Transp Res Rec* 1603:90–98
- Papageorgiou M, Diakaki C, Dinopoulou V, Kotsialos A, Wang Y (2003) Review of road traffic control strategies. In: Proceedings of the IEEE, vol 91, pp 2043–2067
- Papageorgiou M, Wang Y, Kosmatopoulos E, Papamichail I (2007) ALINEA maximizes motorway throughput – an answer to flawed criticism. *Traffic Eng Control* 48(6):271–276
- Payne HJ (1971) Models of freeway traffic and control. In: Bekey GA (ed) Mathematical models of public systems, vol 1. Simulation council, La Jolla
- Payne HJ (1979) FREEFLO: a macroscopic simulation model of freeway traffic. *Transp Res Rec* 772:68–75
- Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Trans Res Rec* 1634:64–69
- Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24:274–287
- Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical restriction versus human overreaction: accident avoidance and two-lane simulations. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) Traffic and granular flow'05. Proceedings of the international workshop on traffic and granular flow. Springer, Berlin, pp 503–508
- Prigogine I, Herman R (1971) Kinetic theory of vehicular traffic. American Elsevier, New York
- Qian Y-S, Feng X, Jun-Wei Zeng J-W (2017) A cellular automata traffic flow model for three-phase theory. *Physica A* 479:509–526

- Rakha H, Tawfik A (2009) Traffic networks: dynamic traffic routing, assignment, and assessment. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9429–9470
- Rehborn H, Klenov SL (2009) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9500–9536
- Rehborn H, Koller M (2014) A study of the influence of severe environmental conditions on common traffic congestion features. *J Adv Transp* 48:1107–1120
- Rehborn H, Palmer J (2008) ASDA/FOTO based on Kerner's three-phase traffic theory in North Rhine-Westphalia and its integration into vehicles. In: *Intelligent Vehicles Symposium*, IEEE, pp 186–191
- Rehborn H, Klenov SL, Palmer J (2011a) An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. *Physica A* 390:4466–4485
- Rehborn H, Klenov SL, Palmer J (2011b) Common traffic congestion features studied in USA, UK, and Germany based on Kerner's three-phase traffic theory. In: *Intelligent Vehicles symposium (IV)*, IEEE, pp 19–24
- Rehborn H, Klenov SL, Koller M (2017) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science + Business Media LLC
- Rempe F, Franek P, Fastenrath U, Bogenberger K (2016) Online freeway traffic estimation with real floating car data. In: *Proceedings of 2016 I.E. 19th international conference on ITS*, pp 1838–1843
- Rempe F, Franek P, Fastenrath U, Bogenberger K (2017) A phase-based smoothing method for accurate traffic speed estimation with floating car data. *Trans Res C* 85:644–663
- Richards PI (1956) Shockwaves on the highway. *Oper Res* 4:42–51
- Saifuzzaman M, Zheng Z (2014) Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transp Res C* 48:379–403
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and granular flow '05*. In: *Proceedings of the international workshop on traffic and granular flow*. Springer, Berlin
- Schadschneider A, Chowdhury D, Nishinari K (2011) *Stochastic transport in complex systems*. Elsevier Science, New York
- Schönhof M, Helbing D (2007) Empirical features of congested traffic states and their implications for traffic modelling. *Transp Sc* 41:135–166
- Schönhof M, Helbing D (2009) *Transp Res B* 43:784–797
- Schreckenberg M, Wolf DE (eds) (1998) *Traffic and granular flow '97*. In: *Proceedings of the international workshop on traffic and granular flow*. Springer, Singapore
- Siebel F, Mauser W (2006) Synchronized flow and wide moving jams from balanced vehicular traffic. *Phys Rev E* 73:066108
- Smaragdis E, Papageorgiou M (2003, 1856) Series of new local ramp metering strategies. *Transp Res Rec*:74–86
- Takayasu M, Takayasu H (1993) Phase transition and 1/f type noise in one dimensional asymmetric particle dynamics. *Fractals* 1:860–866
- Tanga CF, Jiang R, Wu QS (2007) Phase diagram of speed gradient model with an on-ramp. *Physica A* 377:641–650
- Taylor MAP (ed) (2002) *Transportation and traffic theory in the 21st century*. In: *Proceedings of the 15th international symposium on transportation and traffic theory*. Elsevier Science Ltd, Amsterdam
- Tian J-F, Treiber M, Ma SF, Jia B, Zhang WY (2015) Microscopic driving theory with oscillatory congested states: model and empirical verification. *Transp Res B* 71:138–157
- Tian J-F, Jiang R, Jia B, Gao Z-Y, Ma SF (2016a) Empirical analysis and simulation of the concave growth pattern of traffic oscillations. *Transp Res B* 93:338–354
- Tian J-F, Jiang R, Li G, Treiber M, Jia B, Zhu CQ (2016b) Improved 2D intelligent driver model in the framework of three-phase traffic theory simulating synchronized flow and concave growth pattern of traffic oscillations. *Transp Res F* 41:55–65
- Tian J-F, Li G, Treiber M, Jiang R, Jia N, Ma SF (2016c) Cellular automaton model simulating spatiotemporal patterns, phase transitions and concave growth pattern of oscillations in traffic flow. *Transp Res B* 93:560–575
- Tomer E, Safonov L, Havlin S (2000) Presence of many stable nonhomogeneous states in an inertial car-following model. *Phys Rev Lett* 84:382–385
- Treiber M, Kesting A (2013) *Traffic flow dynamics*. Springer, Berlin
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62:1805–1824
- Treiber M, Kesting A, Helbing D (2010) Three-phase traffic theory and two-phase models with a fundamental diagram in the light of empirical stylized facts. *Transp Res B* 44:983–1000
- Treiterer J (1967) Improvement of traffic flow and safety by longitudinal control. *Transp Res* 1:231–251
- Treiterer J (1975) Investigation of traffic dynamics by aerial photogrammetry techniques. Ohio State University technical report PB 246 094, Columbus
- Treiterer J, Myers JA (1974) The hysteresis phenomenon in traffic flow. In: Buckley DJ (ed) *Proceedings of 6th international symposium on transportation and traffic theory*. A.H. & AW Reed, London, pp 13–38
- Treiterer J, Taylor JI (1966) Traffic flow investigations by photogrammetric techniques. *Highway Res Rec* 142:1–12
- Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. *Physica A* 378:475–484
- Wardrop JG (1952) Some theoretical aspects of road traffic research. In: *Proceedings of institute of civil engineering* 1(3):325–362 PART 1. <https://doi.org/10.1680/ipeds.1952.11259>

- Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
- Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
- Wolf DE (1999) Cellular automata for traffic simulations. *Physica A* 263:438–451
- Xiang Z-T, Li Y-J, Chen Y-F, Xiong L (2013) Simulating synchronized traffic flow and wide moving jam based on the brake light rule. *Physica A* 392:5399–5413
- Yang H, Lu J, Hu X-J, Jiang J (2013) A cellular automaton model based on empirical observations of a drivers oscillation behavior reproducing the findings from Kerners three-phase traffic theory. *Physica A* 392:4009–4018
- Zurlinden H (2003) Ganzjahresanalyse des Verkehrsflusses auf Straßen. Schriftenreihe des Lehrstuhls für Verkehrswesen der Ruhr-Universität Bochum, Heft 26, Bochum

Mathematical Probabilistic Approaches to Traffic Breakdown

Boris S. Kerner¹ and Sergey L. Klenov²

¹Physics of Transport and Traffic, University Duisburg-Essen, Duisburg, Germany

²Department of Physics, Moscow Institute of Physics and Technology, Moscow, Russia

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Model of Traffic Breakdown at Highway

Bottleneck

Probabilistic Description of Traffic Breakdown

with Three-Phase Cellular Automaton

(CA) Traffic Flow Model

Probabilistic Description of Traffic Breakdown

Based on Master Equation

Stochastic Highway Capacity of Free Flow at

Bottlenecks

Conclusions

Future Directions

Bibliography

Glossary

Effectual Bottleneck An effectual bottleneck is a highway bottleneck at which an $F \rightarrow S$ transition (traffic breakdown) occurs during many days and years of empirical observations. Because only effectual bottlenecks are considered in the entry, the term *bottleneck* is used below for an effectual bottleneck.

Traffic Breakdown Traffic breakdown is the onset of congested traffic in an initial free traffic flow. In highway traffic, traffic breakdown occurs mostly at effectual highway bottlenecks like on and off-ramps, roadworks,

road gradients, reduction of road lanes, a slow moving vehicle (moving bottleneck), etc. Traffic breakdown results in the emergence of the synchronized flow phase of congested traffic, i.e., traffic breakdown is a phase transition from the free flow traffic phase to synchronized flow traffic phase at a bottleneck ($F \rightarrow S$ transition for short; F – free flow, S – synchronized flow). Thus, the terms *traffic breakdown* and an $F \rightarrow S$ transition are synonyms related to the same phenomenon of the onset of congestion in free flow at the bottleneck.

Nucleation and Probabilistic Features of Traffic Breakdown

Traffic Breakdown Traffic breakdown at a highway bottleneck is an $F \rightarrow S$ transition at the bottleneck that exhibits a nucleation nature: (i) The $F \rightarrow S$ transition occurs in *metastable free flow* at the bottleneck. (ii) At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*. (iii) Onset and dissolution of synchronized flow are accompanied by a *hysteresis* effect. (iv) Traffic breakdown exhibits the following *probabilistic* features: (a) at the same flow rates in some of the realizations (days) traffic breakdown occurs but in other realizations it does not occur; (b) empirical probability of the breakdown is a growing function of the flow rate downstream of the bottleneck; (v) in a realization, the flow rate at which traffic breakdown is observed can be considerably smaller than flow rates at which no traffic breakdown has occurred before. The nucleation nature of traffic breakdown ($F \rightarrow S$ transition) has been proven in empirical traffic data. The empirical nucleation nature of traffic breakdown can be considered an empirical fundamental of transportation science.

Nucleation Model for Traffic Breakdown

In the three-phase traffic theory, it is assumed that the empirical metastability of free flow with respect to an $F \rightarrow S$ transition is governed by an $S \rightarrow F$ instability that exhibits the nucleation nature. The $S \rightarrow F$ instability occurs due to a driver time delay in over-acceleration. It is assumed that within a local speed disturbance

occurring in free flow there is a competition between driver speed adaptation that describes a tendency to synchronized flow and driver over-acceleration that describes a tendency to free flow. The competition between driver speed adaptation and driver over-acceleration is realized within a local speed disturbance in free flow localized at a bottleneck. The disturbance is called a permanent speed (density) local disturbance at the bottleneck. The permanent disturbance can be considered a local vehicle cluster (cluster for short) at the bottleneck. The total number of vehicles within the disturbance can be considered a size of the local vehicle cluster (cluster size for short). Traffic breakdown occurs when driver speed adaptation within a local cluster overcomes on average driver over-acceleration. Such a local cluster can be considered a nucleus for traffic breakdown.

Definition of the Subject and Its Importance

Traffic breakdown is the onset of congested traffic in an initial free traffic flow. Thus, traffic breakdown restricts free flow conditions in vehicular traffic. In observations, traffic breakdown exhibits the nucleation nature. For this reason, the development of the mathematical probabilistic approaches to traffic breakdown is of great importance for all kind of traffic management and control in transportation networks and for other traffic engineering applications.

Introduction

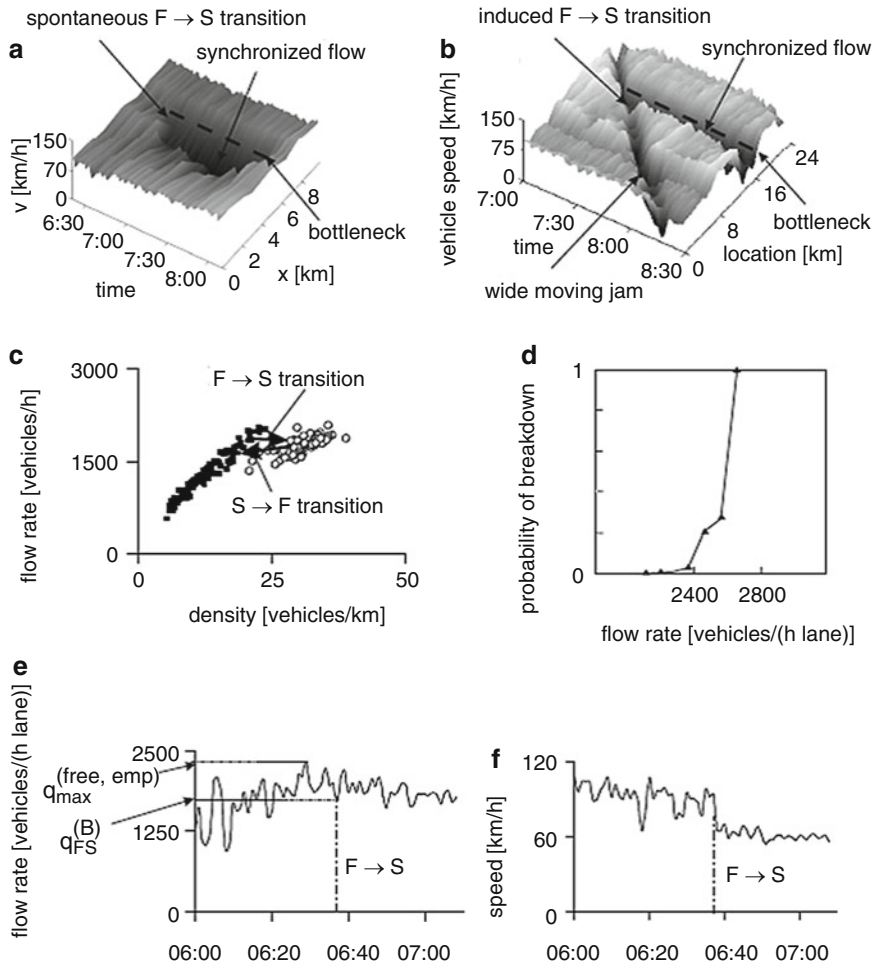
The onset of congestion in an initial free flow is accompanied by a sharp decrease in the average vehicle speed in free flow to a considerably lower speed in congested traffic. This speed breakdown occurs mostly at highway bottlenecks, and it is called the breakdown phenomenon or traffic breakdown (see, e.g., Elefteriadou et al. 1995; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990; Persaud et al. 1998). The traffic breakdown occurs mostly at effectual road bottlenecks like on-

and off-ramps, roadworks, road gradients, reduction of road lanes, etc. An effectual bottleneck is a bottleneck at which traffic breakdown occurs most frequently on many different days.

Beginning from the classical work by Greenshields et al. (1935), a great effort has been made by many scientific groups to understand empirical features of traffic breakdown (see references in Brilon et al. (2005a), Brilon and Zurlinden (2004), Elefteriadou (2014), Elefteriadou et al. (2014, 1995), Hall and Agyemang-Duah (1991), Hall et al. (1992), Lorenz and Elefteriadou (2000), Mahnke et al. (2005), May (1990), and Persaud et al. (1998)).

Traffic breakdown at a highway bottleneck is a local phase transition from free flow (F) to congested traffic whose downstream front is usually fixed at the bottleneck location (Fig. 1a) (see, e.g., Brilon et al. 2005a; Brilon and Zurlinden 2004; Elefteriadou 2014; Elefteriadou et al. 2014, 1995; Hall and Agyemang-Duah 1991; Hall et al. 1992; Lorenz and Elefteriadou 2000; Mahnke et al. 2005; May 1990; Persaud et al. 1998; Hausken and Rehborn 2015; Rehborn and Klenov 2009; Rehborn et al. 2011a, b; entry ► “Traffic Prediction of Congested Patterns” Rehborn and Koller 2014; Rehborn and Palmer 2008; Rempe et al. 2016, 2017). In the three-phase traffic theory (three-phase theory for short), such congested traffic is called synchronized flow (S) (Kerner 2004). In other words, using the terminology of the three-phase theory, traffic breakdown is a transition from free flow to synchronized flow ($F \rightarrow S$ transition) (Kerner 2004).

It has been found (see references in Hall and Agyemang-Duah (1991) and Hall et al. (1992)) that the $F \rightarrow S$ transition, which leads to congested pattern emergence, and the $S \rightarrow F$ transition, which leads to the dissolution of this congested pattern, are accompanied by a well-known *hysteresis effect* and hysteresis loop in the flow-density plane (see, e.g., Elefteriadou 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990): The flow rate in free flow at a bottleneck at which congested pattern emerges is usually larger than the flow rate at which the congested pattern dissolves (Fig. 1a, c).



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 1 Empirical features of traffic breakdown ($F \rightarrow S$ transition): (a, b) spontaneous (a) and induced (b) traffic breakdowns; vehicle speed in space and time; adapted from (Kerner 2004). (c) A well-known (Elefteriadou 2014; Hall and Agyemang-Duah 1991; Hall et al. 1992; May 1990) hysteresis effect associated with (a) (arrows labeled $F \rightarrow S$ and $S \rightarrow F$ show traffic breakdown

and $S \rightarrow F$ transition at bottleneck, respectively). (d) Probability of traffic breakdown as function of flow rate downstream of bottleneck; 10-min average data; adapted from Persaud et al. (1998). (e, f) 1-min averaged flow rate per lane (e) and averaged speed (f) at bottleneck as time-function before and after breakdown (arrows labeled $F \rightarrow S$ show the time instant of the breakdown); adapted from (Kerner 2004)

In 1995, Elefteriadou et al. found that traffic breakdown at a highway bottleneck has a stochastic (probabilistic) behavior (Elefteriadou 2014; Elefteriadou et al. 1995). This means the following: At a given flow rate in free flow at the bottleneck traffic breakdown can occur but it should not necessarily occur. Thus, on one day traffic breakdown occurs; however, on another day at the same flow rates traffic breakdown is not observed.

Persaud et al. (1998) have found that the probability of traffic breakdown at a bottleneck is an increasing function of flow rate (Fig. 1d). Moreover, it turns out (Elefteriadou 2014) that at given traffic parameters (weather, etc.), the maximum flow rate downstream of an on-ramp bottleneck (denoted in Fig. 1e by $q_{\max}^{(\text{free, emp})}$), which was measured on a specific day before congestion occurred, can be larger than the flow rate (denoted in Fig. 1e by $q_{FS}^{(B)}$) at which traffic breakdown occurs (Fig. 1e, f).

There are a huge number of publications devoted to theoretical investigations of traffic breakdown (see references in Brilon et al. (2005a), Brilon and Zurlinden (2004), Elefteriadou (2014), Elefteriadou et al. (2014, 1995), Hall and Agyemang-Duah (1991), Hall et al. (1992), Lorenz and Elefteriadou (2000), Mahnke et al. (2005), May (1990), and Persaud et al. (1998)).

However, the puzzle of empirical features of traffic breakdown has been understood only recently. It was found out that empirical traffic breakdown ($F \rightarrow S$ transition) exhibits the nucleation nature (Kerner 1998b, 2002b). The term *nucleation nature* of traffic breakdown means that traffic breakdown occurs in a metastable free flow with respect to an $F \rightarrow S$ transition. The metastability of free flow is as follows. There can be many speed (density, flow rate) disturbances in free flow. Amplitudes of the disturbances can be very different. When a disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown occurs. Such a disturbance resulting in the breakdown is called a *nucleus* for traffic breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays; as a result, no traffic breakdown occurs.

The empirical nucleation nature of traffic breakdown has been empirically proven by observations of the empirical induced traffic breakdown (induced $F \rightarrow S$ transition) caused by the propagation of a moving spatiotemporal congested pattern to the location of a bottleneck (Fig. 1b) as well as in analyzes of empirical nuclei leading to empirical spontaneous traffic breakdown (Fig. 1a). A detailed empirical study of the nucleation nature of traffic breakdown can be found in Chap. 3 of the book (Kerner 2017a). Thus, traffic breakdown is a local phase transition from a metastable state of the free flow traffic phase to synchronized flow traffic phase ($F \rightarrow S$ transition) (Kerner 2004, 2009c). In other words, the terms traffic breakdown, breakdown phenomenon, speed breakdown, and an $F \rightarrow S$ transition are synonyms that are related to the same phenomenon of the onset of congestion in metastable free flow with respect to an $F \rightarrow S$ transition at a bottleneck.

As explained in reviews (Kerner 2013, 2015, 2016), in the book (Kerner 2017a) as well as in this Encyclopedia (Kerner 2009a; entry ► “Modeling Approaches to Traffic Breakdown”), earlier traffic flow theories and models reviewed in Chowdhury et al. (2000), Cremer (1979), Daganzo (1997), Gartner et al. (2001), Haight (1963), Helbing (2001), Leutzbach (1988), Mahnke et al. (2005, 2009), May (1990), Nagatani (2002), Nagel et al. (2003), Piccoli and Tosin (2009), Rakha et al. (2009), Rakha and Wang (2009), Schadschneider et al. (2011), Treiber and Kesting (2013), Whitham (1974), Wiedemann (1974), and Wolf (1999) cannot explain and predict the empirical nucleation nature of the $F \rightarrow S$ transition (traffic breakdown).

To explain the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway bottlenecks, Kerner introduced three-phase traffic theory (three-phase theory for short) (Kerner 1998b, 1999a, b, 2000a, b, 2001, 2002a, b, 2004). In the three-phase theory, besides the free flow traffic phase there are two traffic phases in congested traffic: synchronized flow and wide moving jam. Thus, there are three traffic phases in this theory (Kerner 2004):

1. Free flow
2. Synchronized flow
3. Wide moving jam

The synchronized flow and wide moving jam phases in congested traffic are defined through spatiotemporal empirical criteria. The definition of the wide moving jam phase [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. The definition of the synchronized flow phase [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

The main reason for the three-phase theory is the explanation of the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway

bottlenecks (Kerner 1998b, 1999a, b, 2000a, b, 2001, 2002a, b, 2004). The three-phase theory explains the empirical nucleation nature of traffic breakdown through a metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck. In the three-phase theory, it is assumed that this free flow metastability occurs through a spatiotemporal competition between driver speed adaptation that describes a tendency to synchronized flow and driver over-acceleration effect that describes a tendency to free flow. It is further assumed that the over-acceleration effect causes an $S \rightarrow F$ instability that governs the nucleation of the $F \rightarrow S$ transition at the bottleneck. Thus, we can make the following conclusions:

- The main prediction of the three-phase theory is the nucleation nature of the $S \rightarrow F$ instability that governs the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck (Kerner 1999a, b, 2000a, b, 2001, 2002a, b, 2004).
- The $S \rightarrow F$ instability and the associated metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck have *no* sense for the classical traffic and transportation theories. Therefore, the three-phase theory is incommensurable with all earlier traffic flow theories and models (for a review, see the book (Kerner 2017a)). The term “incommensurable” has been introduced by Kuhn (2012) to explain a paradigm shift in a scientific field.

The empirical nucleation nature of traffic breakdown as well as modeling approaches to traffic breakdown have already been reviewed in reviews (Kerner 2013, 2015, 2016), in the book (Kerner 2017a) as well as in this Encyclopedia (entries ► “Breakdown in Traffic Networks” and ► “Modeling Approaches to Traffic Breakdown”). In this entry, we discuss mathematical approaches to the probabilistic description of traffic breakdown at highway bottlenecks. The three-phase theory, which explains empirical spatiotemporal probabilistic features of traffic breakdown, is the basis for the probabilistic theory of traffic breakdown reviewed in this entry.

The entry is organized as follows. In section “[Model of Traffic Breakdown at Highway Bottleneck](#),” a qualitative discussion of a nucleation model for traffic breakdown is made. Cellular automaton probabilistic description of traffic breakdown is reviewed in section “[Probabilistic Description of Traffic Breakdown with Three-Phase Cellular Automaton \(CA\) Traffic Flow Model](#).” In section “[Probabilistic Description of Traffic Breakdown Based on Master Equation](#)” based on the nucleation model of section “[Model of Traffic Breakdown at Highway Bottleneck](#),” we consider a probabilistic description of traffic breakdown with master equation. With the use of the mathematical probabilistic description of traffic breakdown with master equation made in section “[Probabilistic Description of Traffic Breakdown Based on Master Equation](#),” in section “[Stochastic Highway Capacity of Free Flow at Bottlenecks](#)” we discuss characteristics of stochastic highway capacity of free flow at a highway bottleneck.

Model of Traffic Breakdown at Highway Bottleneck

In the three-phase theory (Kerner 2004), the possibility of induced and spontaneous traffic breakdowns at the bottleneck (Fig. 1a, b) have been explained by the nucleation nature of breakdown phenomenon as follows (Kerner 2004, 2009c, 2013, 2015, 2016, 2017a). There is a local speed disturbance in free flow that is localized at a bottleneck. Within the localized speed disturbance the speed is lower than outside the disturbance. When driver speed adaptation within the speed disturbance is weaker than driver over-acceleration, then within the disturbance an $S \rightarrow F$ instability occurs. As a result, the average speed within the speed disturbance increases and no traffic breakdown occurs at the bottleneck. On contrary, when the speed adaptation effect within the speed disturbance is stronger than the over-acceleration effect, then traffic breakdown ($F \rightarrow S$ transition) occurs at the bottleneck. A localized speed disturbance with a minimum possible amplitude required for traffic breakdown can be considered a critical speed

disturbance. In its turn, the critical disturbance can be considered a critical nucleus required for traffic breakdown at the bottleneck.

The speed within a critical speed disturbance required for the breakdown should depend on the flow rate downstream of the bottleneck. The smaller the flow rate, the lower the speed within the critical local disturbance required for the breakdown in free flow. In contrast, when the flow rate increases, the speed at the bottleneck within the critical local disturbance required for the breakdown increases. Obviously, the lower the speed within the critical local disturbance, the smaller the probability for the breakdown. This explains the increasing character of the probability of traffic breakdown on the flow rate (Fig. 1d).

We can conclude that when a nucleus appears at a bottleneck, traffic breakdown ($F \rightarrow S$ transition) does occur at the bottleneck. A nucleus with a minimum possible amplitude is a critical speed disturbance that can be considered a critical nucleus required for traffic breakdown. In other words, the probability for traffic breakdown is the probability of a random occurrence of a critical nucleus at the bottleneck. These nucleation features of traffic breakdown ($F \rightarrow S$ transition) predicted in the three-phase theory and found out in empirical traffic data (Kerner 1998b, 1999a, b, 2000a, b, 2001, 2002a, b) (for more details, see reviews and books (Kerner 2004, 2009c, 2013, 2015, 2016, 2017a)) are general ones for metastable states occurring in a diverse variety of physical, chemical, biological, and other complex spatiotemporal systems (see a general theory of metastable systems in the book (Gardiner 1994)).

Traffic flow is a spatially distributed system. Therefore, strictly speaking, a critical speed disturbance (critical nucleus) is characterized by both a critical disturbance amplitude and critical spatial distributions of traffic variables within the critical disturbance (see a microscopic theory of the emergence of a critical speed (density) disturbance required for traffic breakdown in Sects. 5.12 and 5.13 of the book (Kerner 2017a)). However, for the simplicity of the analysis made in this entry, we assume in this entry that spatial distributions of traffic variables within a speed (density) disturbance at a bottleneck are related to critical spatial

distributions. Therefore, under this rough approximation used in this entry a critical speed (density) disturbance at the bottleneck is characterized by the critical disturbance amplitude only.

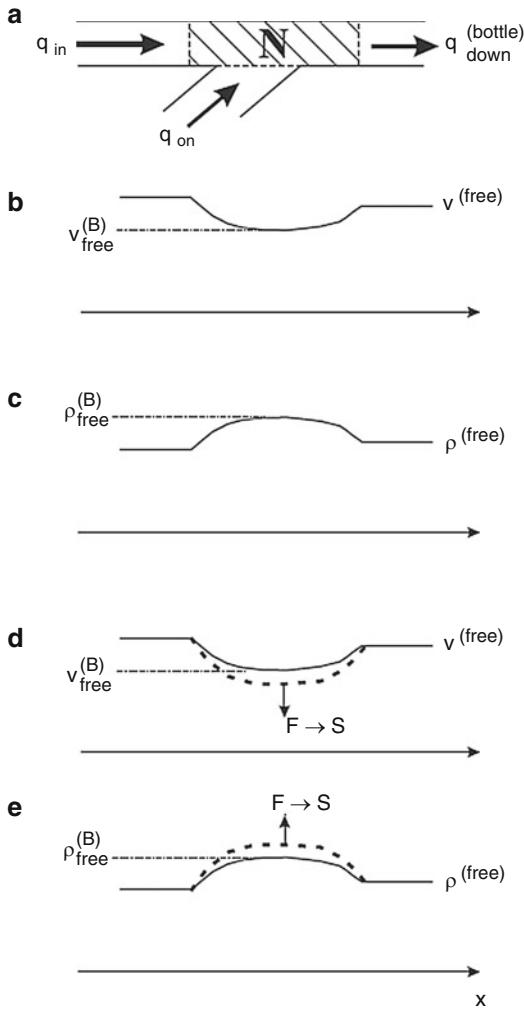
Model of Local Vehicle Cluster at Bottleneck

In accordance with the nucleation model of traffic breakdown based on the three-phase theory briefly discussed above, for a mathematical description of the nucleation of traffic breakdown we assume (Kerner 2000a, 2004; Kerner and Klenov 2005, 2006a, b) that there is a local speed and density disturbance localized at a bottleneck: The speed is lower and density is larger in a neighborhood of the bottleneck than outside the bottleneck. This local speed and density disturbance at the bottleneck can be considered a local *vehicle cluster* in free flow at the bottleneck. One of the characteristics of the cluster is the vehicle cluster size denoted by N . The vehicle cluster size is equal to the *total* number of vehicles within the cluster (Fig. 2a). Traffic breakdown probability within the cluster should be considerably larger than outside the cluster. This should explain why traffic breakdown is observed mostly at bottlenecks.

Even in a hypothetical case in which there were *no random fluctuations* in traffic flow, the vehicle cluster exists at the bottleneck. In this hypothetical case, the permanent local speed and density disturbance at the bottleneck can be considered a *deterministic* vehicle cluster with the cluster size $N = N^{(\text{determ})}$ (a *deterministic* local disturbance) localized at the bottleneck.

Firstly, we consider a highway bottleneck caused by an on-ramp. In the case of the on-ramp bottleneck, the local disturbance in free flow at the bottleneck is caused by two flows, which merge at the bottleneck: (i) An on-ramp inflow with the rate q_{on} . (ii) A flow on the main road upstream of the bottleneck with the rate q_{in} . This flow merging occurs permanent and on the same freeway location (within an on-ramp merging region). For this reason, the vehicle cluster at the bottleneck caused by the local disturbance is on average motionless and permanent (Fig. 2).

To explain features of the vehicle cluster, note that at a given high enough flow rate q_{in} in free



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 2 Qualitative explanations of model of a vehicle cluster in free flow at on-ramp bottleneck: (a) Qualitative schema of on-ramp bottleneck with the vehicle cluster (dashed region) localized at the bottleneck. (b, c) Qualitative spatial speed (b) and density (c) distributions within deterministic vehicle cluster (deterministic local disturbance). (d, e) Dashed curves show qualitative spatial speed (d) and density (e) distributions within the vehicle cluster (local disturbance) at a fixed time instant for the case when fluctuations in free flow lead to an increase in the cluster size N in comparison with the size of a deterministic cluster $N^{(determ)}$. Arrows labeled by “F → S” symbolize traffic breakdown that occurs when a critical cluster appears

flow on the main road upstream of the bottleneck, vehicles that merge from the on-ramp onto the main road force the vehicles on the main road to

decelerate in the vicinity of an on-ramp merging region. This leads to a local decrease in speed and consequently to a local increase in density, i.e., to vehicle cluster occurrence and existence in free flow in the vicinity of the bottleneck.

It must be noted that here and below, the flow rates q_{in} , and q_{on} , as well as all other flow rates used in this entry below, are total flow rates across associated road locations, which are divided by the number of lanes on the main road downstream of the bottleneck. As the flow rates, the cluster size N is also considered per a highway lane: the total vehicle number of vehicles within the cluster is divided by the same number of lanes on the main road downstream of the bottleneck as that used for flow rate unit definition.

As mentioned above, in the hypothetical free flow at a bottleneck in which there were no random fluctuations, a deterministic cluster with the size $N = N^{(determ)}$ occurs at the bottleneck. The speed $v^{(B)}_{free}$ and density $\rho^{(B)}_{free}$ within this deterministic cluster (Fig. 2 (b, c)) correspond to the conditions

$$v^{(B)}_{free} < v^{(free)}, \quad \rho^{(B)}_{free} > \rho^{(free)}, \quad (1)$$

where $v^{(free)}$ and $\rho^{(free)}$ are, respectively, the average vehicle speed and density in homogeneous free flow on the main road downstream of the cluster (Fig. 2). We assume that the flow rates q_{in} and q_{on} are time-independent. For the model of the deterministic cluster in free flow at the bottleneck, the flow rate $q^{(bottle)}_{down}$ downstream of the bottleneck will be denoted by q_{sum} :

$$q_{sum} = q_{in} + q_{on}. \quad (2)$$

For the deterministic cluster in free flow at the bottleneck, the speed and density on the main road satisfy the following condition (Fig. 2):

$$q_{sum} = v^{(free)} \rho^{(free)} = v^{(B)}_{free} \rho^{(B)}_{free}. \quad (3)$$

Condition (3) means that under assumption about the existence of the deterministic cluster the total flow rate does not depend on road location. Condition (3) results from the assumption

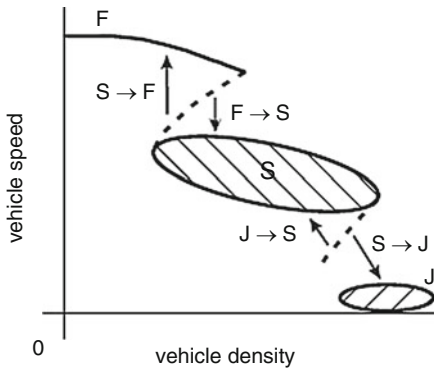
that the size $N^{(\text{determ})}$ of the deterministic cluster at the bottleneck is time-independent and the deterministic cluster is motionless and permanent. In the case, when free flow conditions are at a bottleneck, the flow rate q_{sum} downstream of the bottleneck depends on the bottleneck type. In the cases of off-ramp and merge bottlenecks (a merge bottleneck is a bottleneck caused by a decrease in the number of highway lanes in the flow direction), we get

$$q_{\text{sum}} = q_{\text{in}}. \quad (4)$$

Under condition (4), formulas (1) and (3) are also valid.

Model of Traffic Breakdown Within Vehicle Cluster

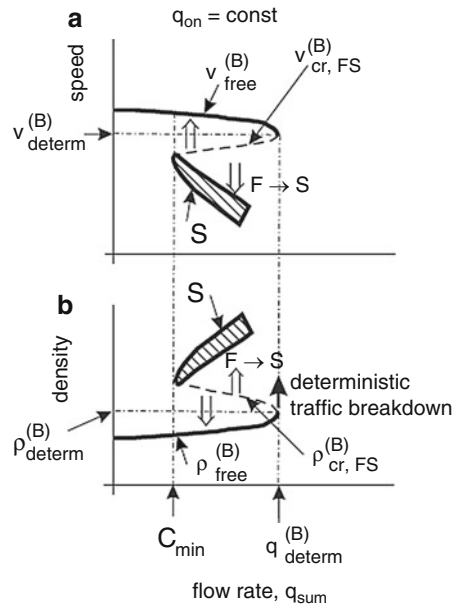
To describe traffic breakdown with the model of vehicle cluster, we use the hypothesis of the three-phase theory about a double Z-speed–density characteristic (2Z–characteristic) (Fig. 3) (Kerner 2004; Kerner and Klenov 2003) (detailed explanations of the 2Z–characteristic can be found in the books (Kerner 2004, 2017a) as well as in the entry ► “Spatiotemporal Features of Traffic Congestion” by Kerner. In this theory, the first Z–speed–density characteristic between states of



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 3 Qualitative double Z speed–density characteristic for phase transitions in the three-phase theory. Arrows labeled by “F → S,” “S → F,” “S → J,” and “J → S” symbolize, respectively, the F → S, S → F, S → J, and J → S transitions. F – free flow traffic phase, S – synchronized flow traffic phase, J – wide moving jam traffic phase (Adapted from Kerner (2004, 2009a))

free flow (F) and states of synchronized flow (S) explains traffic breakdown (arrow labeled by “F → S” in Fig. 3). In the model of traffic breakdown within the vehicle cluster at a bottleneck considered in this entry only the Z–speed–density characteristic between states F and S will be further used (detailed explanations of the Z-characteristic between states F and S can be found in the book (Kerner 2017a) as well as in the entry ► “Modeling Approaches to Traffic Breakdown” by Kerner in this Encyclopedia).

In accordance with the Z-characteristic between states F and S (Fig. 3), at a given value of the on-ramp inflow rate q_{on} there should be nonmonotonous dependence of the vehicle speed on the flow rate q_{in} at the location of vehicle cluster (Fig. 4a). This speed–flow characteristic leads to a S-shaped density dependence on the flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ (Fig. 4b).



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 4 Qualitative shapes of speed (a) and density (b) dependencies on the flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ at a given value of the on-ramp inflow rate q_{on} associated with location of vehicle cluster at on-ramp bottleneck. Arrows labeled by “F → S” and “S → F” symbolize, respectively, the F → S and S → F transitions. F – free flow traffic phase, S – synchronized flow traffic phase (Adapted from Kerner and Klenov (2006b))

When q_{on} is a given value and the flow rate q_{in} increases, the vehicle speed in free flow within the deterministic cluster $v_{free}^{(B)}$ decreases and in accordance with Eq. 3 the associated density $\rho_{free}^{(B)}$ increases. However, this increase in the density is limited by some critical density $\rho_{free}^{(B)} = \rho_{determin}^{(B)}$ (Fig. 4b) within the deterministic cluster associated with a critical flow rate

$$q_{sum} = q_{determin}^{(B)}. \quad (5)$$

After this critical deterministic cluster is reached, as can be seen from Fig. 4b, the further increase in the flow rate $q_{in} + q_{on}$ should lead to *deterministic traffic breakdown* at the bottleneck causing spontaneous synchronized flow emergence at the bottleneck (arrow labeled by “deterministic traffic breakdown” in Fig. 4b). The critical deterministic cluster can be considered a *deterministic nucleus* for traffic breakdown. After the critical deterministic cluster is reached, deterministic traffic breakdown occurs at the bottleneck even if there were no random fluctuations in free flow at the bottleneck. For this reason, we can consider the flow rate $q_{sum} = q_{determin}^{(B)}$ Eq. 5 as the maximum highway capacity of free flow denoted by C_{max} :

$$C_{max} = q_{determin}^{(B)}. \quad (6)$$

The sense of the maximum highway capacity C_{max} given by formula (6) is as follows. If there were no random fluctuations in free flow, at

$$q_{sum} \geq C_{max} \quad (7)$$

due to the deterministic traffic breakdown traffic breakdown does occur at the bottleneck.

Due to fluctuations in real free flow, spontaneous traffic breakdown ($F \rightarrow S$ transition) can occur within the flow rate range

$$C_{min} \leq q_{sum} < q_{determin}^{(B)}, \quad (8)$$

i.e., before the deterministic nucleus for traffic breakdown is reached (arrows labeled by “ $F \rightarrow S$ ” in Fig. 4). In Eq. 8, the characteristic

flow rate $q_{sum} = C_{min}$ is a minimum highway capacity of free flow at the bottleneck, which is considerably less than $q_{determin}^{(B)}$: At

$$q_{sum} < C_{min}, \quad (9)$$

no traffic breakdown is possible at the bottleneck.

Under condition (8), free flow at a bottleneck is in a metastable state with respect to the $F \rightarrow S$ transition. Using formula (6), the flow rate range Eq. 8 of free flow metastability with respect to the $F \rightarrow S$ transition can be rewritten as follows:

$$C_{min} \leq q_{sum} < C_{max}. \quad (10)$$

It should be noted that in section “Definition of Stochastic Highway Capacity” we will continue a consideration of condition (10) in the relation with the definition of stochastic highway capacity of free flow at the bottleneck.

To explain the model of traffic breakdown under consideration, we note that due to fluctuations in real free flow at the bottleneck, the cluster size N can exceed the deterministic cluster size $N^{(determin)}$. In this case, the vehicle density increases (Fig. 2e) and speed decreases (Fig. 2d) within the cluster in free flow at the bottleneck in comparison with the density (Fig. 2c) and speed (Fig. 2b) within the deterministic cluster. In other words, due to fluctuations in real free flow, the speed and density within the vehicle cluster at the bottleneck and, therefore, the cluster size N can differ considerably from, respectively, the values $v_{free}^{(B)}$, $\rho_{free}^{(B)}$, and $N^{(determin)}$ for the deterministic cluster. The vehicle cluster with the size N can be considered consisting of two components: (a) the deterministic cluster with the size $N^{(determin)}$ and (b) a random cluster component with the size $N - N^{(determin)}$.

In the model, there are a speed within a critical cluster denoted by $v_{cr,FS}^{(B)}$ and the associated density $\rho_{cr,FS}^{(B)}$ within the critical cluster, which depend on the flow rate (Fig. 4). When within the cluster the speed is equal to or lower than $v_{cr,FS}^{(B)}$, respectively, the density is equal to or larger than $\rho_{cr,FS}^{(B)}$, then traffic breakdown occurs spontaneously. Thus, when the cluster size N exceeds some critical cluster size associated with the density $\rho_{cr,FS}^{(B)}$

within the cluster (Fig. 4b), then traffic breakdown occurs. The cluster with the critical size can be considered a critical vehicle cluster or a critical nucleus for traffic breakdown at the bottleneck. Otherwise, if the cluster size N is smaller than the critical one, the cluster decays toward the initial deterministic cluster with the cluster size $N^{(\text{determ})}$. Thus, under condition (8) fluctuations in the cluster size in the neighborhood of $N = N^{(\text{determ})}$ are responsible for traffic breakdown (a more detailed explanation of the effect of fluctuations appears in section “Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory”).

Probabilistic Description of Traffic Breakdown with Three-Phase Cellular Automaton (CA) Traffic Flow Model

The KKW (Kerner-Klenov-Wolf) cellular automaton (CA) model is the first CA traffic flow model developed in the framework of the three-phase theory, which can show and predict the empirical nucleation nature of traffic breakdown through the metastability of free flow with respect to an $F \rightarrow S$ transition at a highway bottleneck (Kerner et al. 2002). Simulations of this model show that in accordance with the nucleation model of traffic breakdown considered above there is a vehicle cluster, i.e., a local density (and speed) disturbance in free flow, which is localized at the bottleneck. In the KKW-model, as in the nucleation model of section “Model of Traffic Breakdown at Highway Bottleneck,” traffic breakdown in free flow at the bottleneck occurs mostly at the location of the vehicle cluster.

As mentioned, in the KKW CA model traffic breakdown exhibits the nucleation nature: Traffic breakdown is an $F \rightarrow S$ transition that occurs in a metastable free flow with respect to the $F \rightarrow S$ transition. It has been found (Kerner et al. 2002) that traffic breakdown in the KKW CA model shows all main features of empirical traffic breakdown at a highway bottleneck mentioned in section “Introduction” (Fig. 1):

- (i) In a finite flow rate range, there can be either spontaneous or induced traffic breakdown ($F \rightarrow S$ transition) at the same bottleneck.

- (ii) There is a hysteresis effect associated with traffic breakdown ($F \rightarrow S$ transition) and a return $S \rightarrow F$ transition.
- (iii) The breakdown probability is an increasing function of the flow rate q_{sum} (see section “Time Delay and Probability of Traffic Breakdown”).

For these reasons, we briefly discuss this traffic flow model as well as a probabilistic theory of traffic breakdown associated with the model.

KKW Three-Phase Traffic Flow CA Model

The KKW CA model reads as follows (Kerner et al. 2002):

$$v_{n+1} = \max(0, \min(\tilde{v}_{n+1} + a\tau\eta_n, v_n + a\tau, v_{\text{free}}, v_{s,n})), \quad (11)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (12)$$

$$\tilde{v}_{n+1} = \max(0, \min(v_{\text{free}}, v_{c,n}, v_{s,n})), \quad (13)$$

$$v_{c,n} = \begin{cases} v_n + a\tau & \text{at } g_n > G_n, \\ v_n + a\tau \text{sign}(v_{\ell,n} - v_n) & \text{at } g_n \leq G_n, \end{cases} \quad (14)$$

where index n corresponds to the discrete time $t = n\tau$, $n = 0, 1, 2, \dots$; τ is time step, v_n is the speed at time step n , $\tilde{v}_n$ is the speed at time step n without model speed fluctuations, random speed fluctuations $a\tau\eta_n$ explained below, a the maximum vehicle acceleration that is a constant value, v_{free} is the maximum speed in free flow, $g_n = x_{\ell,n} - x_n - d$ is a space gap between vehicles at time step n , x_n is the vehicle co-ordinate, the subscript ℓ marks functions and values related to the preceding vehicle, G_n is a synchronization space gap, $v_{s,n}$ is a safe speed taken through formula $v_{s,n} = g_n/\tau$ that is the same as in the Nagel-Schreckenberg (NaSch) CA model (Nagel and Schreckenberg 1992), the safe speed $v_{s,n}$ determines the safe space gap $g_{\text{safe},n}$ from the equation $v_n = v_{s,n}(g_{\text{safe},n})$; all vehicles have the same length d , which includes the minimum space gap between vehicles within a wide moving jam. A synchronization space gap G_n is taken as a speed function:

$$G(v_n) = kv_n\tau, \quad (15)$$

where k is constant.

Driver Behavioral Assumptions of KKW CA Model

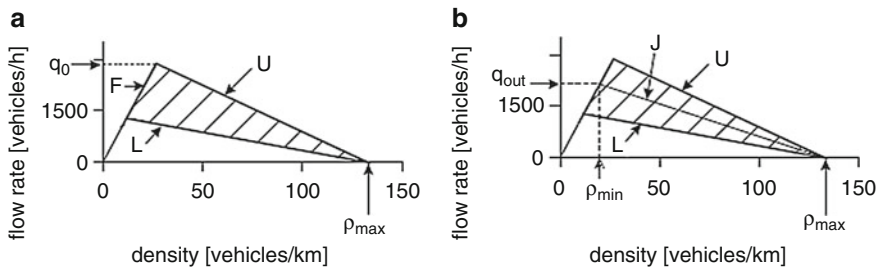
Basic driver behavioral assumptions of the KKW CA model are similar to those of the our stochastic three-phase traffic flow model (Kerner and Klenov 2002, 2003, 2009) discussed in Kerner (2009a, 2017a; entry ► “Modeling Approaches to Traffic Breakdown”):

- (i) *A hypothesis of the three-phase theory about a 2D-region of steady states of synchronized flow* (Kerner 1998a, b, 1999a, b, 2004). In synchronized flow, a driver accepts a range of different hypothetical steady state speeds at the same space gap to the preceding vehicle. Respectively, at a given time-independent speed in synchronized flow, a driver accepts an arbitrary space gap to the preceding vehicle within a limited space gap range. This means that hypothetical steady model states of synchronized flow constitute a 2D-region in the flow–density plane (Fig. 5a). The boundaries of this 2D-region F , L , and U are respectively associated with free flow, a synchronization space gap, and a safe space gap determined through a safe speed. The 2D-region of the steady states is associated with the following driver behavioral assumption. In synchronized flow, even when the speed difference between the vehicle and the preceding vehicle is very small, a driver is able to recognize whether the space gap is increasing or decreasing over time. This is true for each synchronized flow

speed within a finite space-gap range. Therefore, we can assume that a given vehicle steady speed can be related to the infinite number of space gaps; respectively, a given space gap between vehicles can be related to the infinite number of steady speeds.

- (ii) *Line J and 2D region of steady states* (Kerner 1998a, b, 1999a, b, 2004). In the model, the line J in the flow–density plane, which represents in this plane the motion of the downstream jam front (the line J is explained in the book (Kerner 2004) as well as in the entry ► “Spatiotemporal Features of Traffic Congestion” by Kerner in this Encyclopedia), is between the boundaries L and U (Fig. 5b). Thus, at a given steady speed, a driver behavioral assumption is that the space gap in synchronized flow associated with the line J is larger than the safe one and it is smaller than the synchronization gap. The line J divides the 2D region of steady states of synchronized flow onto two classes (Kerner 1998a, b, 1999a, b, 2004):
 1. States that are on and above the line J , which are metastable states with respect to wide moving jam formation.
 2. States that are below the line J , which are stable states with respect to wide moving jam formation.
- (iii) *Speed adaptation effect in synchronized flow* (Kerner 1998a, b, 1999a, b, 2004). The speed adaptation effect takes place, when the vehicle cannot pass the preceding vehicle, within the space gap range:

$$g_{\text{safe},n} \leq g_n \leq G_n. \quad (16)$$



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 5 Steady states of the KKW CA model and the line J . $\rho_{\max}$ is the vehicle density within a wide moving jam (Adapted from Kerner et al. (2002))

Under condition (16), as follows from Eq. 14, the vehicle tends to adjust its speed to the preceding vehicle without caring, what the precise space gap is, as long as it is safe, i.e., there is no desired (or optimal) space gap in synchronized flow.

- (iv) *Over-acceleration effect* (Kerner 1999a, b, 2004). The driver over-acceleration effect is driver maneuver leading to a higher speed from initial car-following at a lower speed occurring under condition (16) (for more details, see (Kerner 2017a; entry ► “Modeling Approaches to Traffic Breakdown”)). In the three-phase theory, it is assumed that the probability of driver over-acceleration exhibits a discontinuity as a function of the vehicle density: At the same density, the probability of driver over-acceleration is larger in free flow than that in synchronized flow. A competition between the speed adaptation (item (iii)) and over-acceleration effects simulates the metastability of free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at road bottlenecks. In the KKW CA model, driver over-acceleration is simulated as a collective effect, which occurs on average in traffic flow, through the use of random vehicle acceleration. In KKW CA model, both random acceleration and random deceleration are simulated by adding the same random term $a\tau\eta_n$ to the vehicle speed in Eq. 11, where

$$\eta_n = \begin{cases} -1 & \text{if } r < p_b, \\ 1 & \text{if } p_b \leq p_b + p_a, \\ 0 & \text{otherwise,} \end{cases} \quad (17)$$

p_a is probability of random acceleration, p_b is probability of random deceleration, $p_a + p_b \leq 1$, $r = \text{rand}(0, 1)$ is a random number uniformly distributed between 0 and 1. The random vehicle acceleration is described by the second line in Eq. 17.

- (v) *Pinch effect in synchronized flow*. The pinch effect in synchronized flow is an increase density of synchronized flow with the

following spontaneous emerging of growing narrow moving jams in the dense synchronized flow (called as an $S \rightarrow J$ instability) (for more details, see (Kerner 2004, 2009b, c; entry ► “Spatiotemporal Features of Traffic Congestion”)). Moving in synchronized flow, a driver comes on average closer to the preceding vehicle over time that should explain the pinch effect. This effect is simulated in the KKW CA model through the increase in random vehicle acceleration at low vehicle speeds. For this purpose, the probability p_a is chosen to be a decreasing speed function

$$p_a(v_n) = \begin{cases} p_{a1} & \text{if } v_n < v_p \\ p_{a2} & \text{if } v_n \geq v_p, \end{cases} \quad (18)$$

where p_{a1} , p_{a2} , and v_p are constants; $p_{a1} > p_{a2}$.

- (vi) *Over-deceleration effect*. The over-deceleration effect introduced by Herman, Gazis, Montroll, Potts, Rothery, and Chandler in 1958–1961 is as follows (Chandler et al. 1958; Gazis 2002; Gazis et al. 1959, 1961; Herman et al. 1959) (for more details, see (Kerner 2009a, 2017a)): If a vehicle begins to decelerate unexpectedly, then due to the finite driver reaction time the following vehicle starts deceleration with a delay. When the driver reaction time is long enough, the driver of the following vehicle decelerates stronger than it is needed to avoid collisions. As a result, the speed of the following vehicle becomes lower than the speed of the preceding vehicle. If this over-deceleration effect is realized for the following drivers, a wave of vehicle speed reduction should appear and grow over time. We call the traffic flow instability due to the over-deceleration effect (driver reaction time) as the classical traffic flow instability of Herman, Gazis, Montroll, Potts, Rothery, and Chandler. In the KKW CA model, a competition between this well-known over-deceleration associated with driver reaction time and the speed adaptation effect determines moving jam emergence in synchronized flow ($S \rightarrow J$ instability).

A consideration of the $S \rightarrow J$ instability is out of scope of this entry (see the book (Kerner 2004) as well as the entry ► “Spatiotemporal Features of Traffic Congestion” by Kerner in this Encyclopedia). As in the NaSch CA model (Nagel and Schreckenberg 1992), the over-deceleration is also simulated as a collective effect through the use of random fluctuations in vehicle deceleration. The random vehicle deceleration is described by the first line in formula (17).

- (vii) *Driver time delay in acceleration.* In the model, this well-known effect should describe driver delay in acceleration at the downstream front of synchronized flow or wide moving jam after the preceding vehicle has begun to accelerate (for more details, see Kerner 2009a, 2017a). A driver time delay in acceleration is simulated as a collective effect through the use of a random vehicle deceleration (Eq. 17). Indeed, a vehicle that has to accelerate at the current time step can retain its speed due to applying of random vehicle deceleration. In the KKW CA model, as in a version of the NaSch CA model made in Barlović et al. 1998) the slow-and-start rules are used to describe a longer driver time delay in acceleration when the vehicles begins to accelerate from a standstill:

$$p_b(v_n) = \begin{cases} p_0 & \text{if } v_n = 0 \\ p & \text{if } v_n > 0, \end{cases} \quad (19)$$

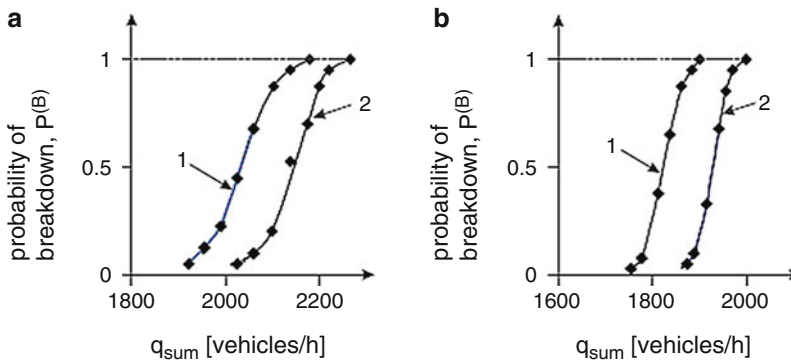
where $p, p_0 > p$ are constants (Barlović et al. 1998); consequently, the mean time in vehicle acceleration from a vehicle standstill is

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p_0}. \quad (20)$$

Thus, the basis of the KKW three-phase CA model are driver behavioral assumptions made in the three-phase theory (items (i)–(v)). In addition, over-deceleration (item vi) and driver time delay in acceleration (item vii) introduced in the NaSch CA model (Barlović et al. 1998; Nagel and Schreckenberg 1992) (see review (Kerner 2009a) in this Encyclopedia) have also been incorporated.

Time Delay and Probability of Traffic Breakdown

Simulations of the KKW CA model (Kerner et al. 2002) show that at a given flow rate in free flow on the main road upstream of an on-ramp bottleneck q_{in} , free flow is metastable with respect to a local $F \rightarrow S$ transition, which can occur spontaneously at the bottleneck, when the flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ in free flow downstream of the bottleneck is satisfied condition (8).



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 6 Flow rate dependencies of the probability $P^{(B)}(q_{\text{sum}})$ that an $F \rightarrow S$ transition occurs at the bottleneck within $T_{\text{ob}} = 30$ min (curve 1) or already within $T_{\text{ob}} = 15$ min (curve 2), after the on-ramp inflow was

switched on. $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$. Results are shown for two different flow rates to the on-ramp, $q_{\text{on}} = 60$ vehicles/h in (a) and $q_{\text{on}} = 200$ vehicles/h in (b) (Adapted from Kerner et al. (2002))

Theoretical probability for an $F \rightarrow S$ transition at the bottleneck (Fig. 6) shows the same features as empirical probability for the $F \rightarrow S$ transition (Persaud et al. 1998) (for detailed explanations, see Kerner (2009a, 2017a)). To study the $F \rightarrow S$ transition probability (Fig. 6), a large number of realizations (runs) have been calculated for each given set of the flow rates $q_{\text{sum}} = q_{\text{on}} + q_{\text{in}}$ and q_{on} during a given time interval of the observing traffic flow T_{ob} . In different simulation runs only initial conditions (at $t = 0$) for random model fluctuations have been different. At the beginning of each run there was free flow at the on-ramp bottleneck. For each run it was checked, whether the $F \rightarrow S$ transition at the on-ramp bottleneck occurred during the time interval T_{ob} or not. The result of these simulations is the number of realizations n_r where the $F \rightarrow S$ transition at the on-ramp had occurred in comparison with the number of all realizations N_r . Then the probability for an $F \rightarrow S$ transition is $P^{(B)} = n_r/N_r$.

The flow rate q_{sum} was changed through the change in q_{in} and the procedure with all realizations was repeated at the same chosen flow rate to the on-ramp q_{on} . The flow rate q_{sum} at which the $F \rightarrow S$ transition at the on-ramp occurred in all realization is related to the probability of the $F \rightarrow S$ transition $P^{(B)} = 1$. We found that lower flow rates q_{sum} correspond to $P^{(B)} < 1$. As expected, we found an increase of the probability $P^{(B)}$ as a function of the flow rate q_{sum} at a given flow rate to the on-ramp q_{on} ; the related function found numerically can be approximated by the formula (Kerner et al. 2002)

$$P^{(B)}(q_{\text{sum}}) = \frac{1}{1 + \exp[\beta(q_p - q_{\text{sum}})]}, \quad (21)$$

where β and q_p are parameters that are functions of q_{on} and T_{ob} .

Traffic breakdown is characterized by a random time delay $T^{(B)}$: In different simulation realizations (runs), at the same set of the flow rates q_{on} and q_{in} usually different $T^{(B)}$ have been found (see, e.g., Fig. 5.11 of the book Kerner (2017a)). The smaller q_{on} and/or q_{in} are, the longer is the mean time delay $T^{(B, \text{mean})}$ of traffic breakdown.

Probabilistic Description of Traffic Breakdown Based on Master Equation

The nucleation model for traffic breakdown of section “[Model of Traffic Breakdown at Highway Bottleneck](#),” which is confirmed by numerical simulations of the KKW CA model discussed above, can be used for an analytical study of the nucleation rate, breakdown probability, and mean time delay of traffic breakdown at a bottleneck based on the well-known master equation approach for probabilistic description of phase transitions in complex metastable systems of diverse nature (Gardiner 1994).

Master Equation and One-Step Processes

Traffic flow variables (speed, density, occupancy, flow rate) within the vehicle cluster localized at a bottleneck (section “[Model of Traffic Breakdown at Highway Bottleneck](#)”) are time dependent random variables. A time dependent random variable $X(t)$ can be characterized by a set of joint probabilities (Gardiner 1994)

$$p(x_1, t_1; x_2, t_2; \dots; x_n, t_n), \quad n = 1, 2, \dots, \quad (22)$$

where $x_1, x_2, \dots, x_n$ are random values of variable X at time instants $t_1 \leq t_2 \leq \dots \leq t_n$. The joint probabilities $p(x_1, t_1; \dots; x_n, t_n)$ and $p(x_1, t_1; \dots; x_{n-1}, t_{n-1})$ are connected to each other by a conditional probability

$$p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1}), \quad n = 2, 3, \dots, \quad (23)$$

as follows (Gardiner 1994)

$$p(x_1, t_1; \dots; x_n, t_n) = p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1}) p(x_1, t_1; \dots; x_{n-1}, t_{n-1}). \quad (24)$$

The basic assumption of *Markovian* stochastic process is that the conditional probability $p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1})$ depends on the value of random variable X at the last previous time moment t_{n-1} only (Gardiner 1994)

$$p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1}) = p(x_n, t_n | x_{n-1}, t_{n-1}). \quad (25)$$

In this case, according to Eq. 24 all joint probabilities $p(x_1, t_1; \dots; x_n, t_n)$ Eq. 22 at $n > 1$ can be expressed in terms of the conditional probability (25) and the probability $p(x_1, t_1)$ as

$$p(x_1, t_1; \dots; x_n, t_n) = p(x_1, t_1) \prod_{k=2}^n p(x_k, t_k | x_{k-1}, t_{k-1}). \quad (26)$$

In the nucleation model of section “[Model of Traffic Breakdown at Highway Bottleneck](#),” it is assumed that there is a vehicle cluster localized at a bottleneck and this cluster exists in free flow even if there were no fluctuations in the flow. In the model, the cluster dynamics describes traffic breakdown ($F \rightarrow S$ transition) at the bottleneck. For an analytical study of the nucleation model of section “[Model of Traffic Breakdown at Highway Bottleneck](#),” we will further make the following additional assumptions: (i) The cluster dynamics is associated with a Markovian stochastic process; (ii) the random variable X of this process is associated with the total number of vehicles in the cluster N only (see section “[Outflow Rate from Cluster](#)” for more detail). Then X is a discrete variable. In this case, the Chapman–Kolmogorov equation for the conditional probability $p(x, t | x', t')$ reads (Gardiner 1994):

$$p(x, t | x', t') = \sum_{x''} p(x, t | x'', t'') p(x'', t'' | x', t'), \quad (27)$$

where $t \geq t'' \geq t'$, the summation is taken over all possible values of x'' . In turn, the conditional probability $p(x, t | x', t')$ determines the time evolution of the probability $p(x, t)$ as

$$p(x, t) = \sum_{x'} p(x, t | x', t') p(x', t'). \quad (28)$$

The differential equation for the conditional probability $p(x, t | x', t')$ that follows from Eq. 27 can be written as (Gardiner 1994)

$$\frac{\partial p(x, t | x', t')}{\partial t} = \sum_{y \neq x} [W(x, y, t) p(y, t | x', t') - W(y, x, t) p(x, t | x', t')], \quad (29)$$

where

$$W(x, y, t) = \lim_{\Delta t \rightarrow 0} p(x, t + \Delta t | y, t) / \Delta t \quad \text{at } x \neq y \quad (30)$$

is the transition rate, i.e., the probability per unit time for the transition from y to $x \neq y$ at time t . The Eq. 29 is usually called *master equation*. Multiplying Eq. 29 by $p(x', t')$ and using Eq. 28, one can write (29) as the equation for probability $p(x, t)$ (Gardiner 1994)

$$\frac{\partial p(x, t)}{\partial t} = \sum_{y \neq x} [W(x, y, t) p(y, t) - W(y, x, t) p(x, t)]. \quad (31)$$

In the model of traffic breakdown (sections “[Model of Traffic Breakdown at Highway Bottleneck](#)” and “[Outflow Rate from Cluster](#)”), we will make an additional assumption that at each time instant the total number of vehicles within the cluster N can be changed by one vehicle only. This assumption is associated with a *one-step process*, also called *birth-and-death process*, when transitions between neighboring states $X = x$ and $X = x \pm 1$ can only occur (Gardiner 1994). The transition rate $W(x, y, t)$ for one-step processes is (Gardiner 1994)

$$W(x, y, t) = w_+(y) \delta_{x, y+1} + w_-(y) \delta_{x, y-1}, \quad (32)$$

where $w_+(y)$ is the rate of transition from y to $x = y + 1$; $w_-(y)$ is the rate of transition from y to $x = y - 1$; $\delta_{x, y}$ is Kroneker symbol.

Substituting (Eq. 32) into the master Eq. 31, one finds the master equation for one-step processes (Gardiner 1994)

$$\frac{\partial p(x, t)}{\partial t} = w_+(x-1) p(x-1, t) + w_-(x+1) p(x+1, t) - [w_+(x) + w_-(x)] p(x, t). \quad (33)$$

For the master Eq. 33, the mean first passage time T for transition from an initial point $x = a$ to a

finite point $x = b$ is given by formula (Gardiner 1994)

$$T = \sum_{x=a}^b \left[(w_+(x)p_s(x))^{-1} \sum_{y=0}^x p_s(y) \right], \quad (34)$$

where $p_s(x)$ is a steady solution of Eq. 33:

$$p_s(x) = p_s(0) \prod_{y=1}^x \frac{w_+(y-1)}{w_-(y)} \quad \text{at } x > 0, \quad (35)$$

the reflecting boundary condition

$$w_-(0) = 0 \quad (36)$$

is assumed to be satisfied at the point $x = 0$.

If the points a and b are separated by a potential barrier and therefore $p_s^{-1}(x)$ has a strong maximum at a point $x = c$ between a and b , the formula (34) is reduced to (Gardiner 1994):

$$T = \frac{1}{w_+(c)} \sum_{x=0}^c p_s(x) \sum_{x=a}^b p_s^{-1}(x). \quad (37)$$

We will use this formula for the evaluation of the mean time delay for traffic breakdown $T_{FS}^{(B, \text{mean})}$ in section “Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks.”

Critical Discussion of Earlier Probabilistic Traffic Breakdown Models

It must be noted that models for vehicle cluster nucleation in traffic flow based on the master Eq. 33, which should describe traffic breakdown, have firstly been introduced by Mahnke et al. and Kühne et al. (Kühne et al. 2002, 2004; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997) (see reviews (Mahnke et al. 2005, 2009)). However, in Kerner and Klenov (2006a, b) we have explained in details that and why the models of references (Kühne et al. 2002, 2004; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005) do not explain the empirical nucleation nature of traffic breakdown associated with the metastability of free flow with respect to an

$F \rightarrow S$ transition. In other words, the models of Mahnke et al. and Kühne et al. (Kühne et al. 2002; Kühne et al. 2004; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005) cannot describe the probabilistic traffic breakdown (probabilistic $F \rightarrow S$ transition) in real traffic flow. In particular, the nucleation models of references (Kühne et al. 2002; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005, 2009) describe the nucleation of a wide moving jam in free flow ($F \rightarrow J$ transition). This is in contrast with the empirical evidence that traffic breakdown is the $F \rightarrow S$ transition, as it is observed in all measured traffic data.

Briefly, the most crucial differences of the nucleation models of references (Kühne et al. 2002, 2004; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005) with the model of traffic breakdown of section “Model of Traffic Breakdown at Highway Bottleneck” (Kerner 2000a, 2004; Kerner and Klenov 2005, 2006a, b) are as follows.

The basic assumption of the nucleation models of references (Kühne et al. 2002, 2004; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005) is that there are *no* vehicle clusters on a road without fluctuations. This basic assumption is made both for a homogeneous road (Kühne et al. 2002; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005) and for a road with a bottleneck (Kühne et al. 2004; Mahnke et al. 2005). It is assumed that for vehicle cluster emergence firstly a random precluster should occur from fluctuations in free flow. This precluster should forego subsequent cluster evolution toward a critical cluster (nucleus) whose growth should lead to traffic breakdown. In other words, it is suggested that without random precluster the number of vehicles within the cluster, i.e., the cluster size is exactly equal to *zero* and, consequently, without random fluctuations no traffic breakdown is possible.

In contrast with the nucleation models of references (Kühne et al. 2002, 2004; Mahnke and Kaupužs 1999; Mahnke and Pieret 1997; Mahnke et al. 2005, 2009), in the nucleation model of section “Model of Traffic Breakdown at Highway Bottleneck” (Kerner and Klenov 2005, 2006a, b)

the cluster size N is equal to the *total* number of vehicles within the cluster. Therefore, even if there were no random fluctuations in free flow, rather than the cluster size is equal to zero, the cluster size $N = N^{(\text{determ})}$, i.e., this is equal to the size of the deterministic cluster (Fig. 7) and, consequently, the deterministic traffic breakdown is possible (Kerner and Klenov 2005, 2006a, b). The nucleation model of section “[Model of Traffic Breakdown at Highway Bottleneck](#)” results from the model assumption that an initial free flow at a bottleneck is nonhomogeneous, i.e., a vehicle cluster exists in free flow even if random fluctuations in traffic flow were not taken into account.

In the model of section “[Model of Traffic Breakdown at Highway Bottleneck](#)” (Kerner and Klenov 2005, 2006a, b), due to fluctuations in real free flow, the cluster size N changes in a neighborhood of $N = N^{(\text{determ})}$ (Fig. 7). These fluctuations in the cluster size lead to traffic breakdown before the condition for the deterministic breakdown is satisfied. Below we will see that this nucleation model explains the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition).

Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory

Outflow Rate from Cluster

An important characteristic of the nucleation model of section “[Model of Traffic Breakdown](#)

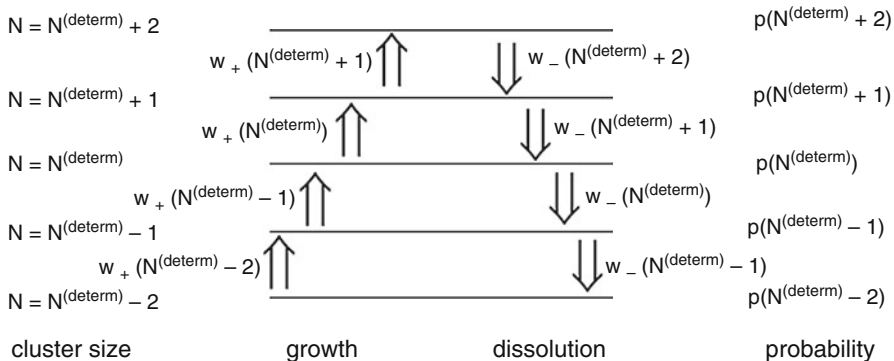
at Highway Bottleneck” is the outflow rate from the vehicle cluster $q_{\text{down}}^{(\text{bottle})}$ (Figs. 2a and 8). In contrast with the on-ramp and merge bottlenecks for which there is only one outflow from the cluster, for the off-ramp bottleneck, there are two different outflows from the cluster, which are related to the flow rate on the main road downstream of the merging region of the off-ramp denoted by $q_{\text{down}}^{(\text{main})}$ and to the flow rate to the off-ramp road denoted by q_{off} . This means that for the off-ramp bottleneck (Fig. 8b)

$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{main})} + q_{\text{off}}. \quad (38)$$

In the nucleation model for traffic breakdown discussed in section “[Model of Traffic Breakdown at Highway Bottleneck](#),” we make further an assumption that the outflow rate from the cluster depends on the total number of vehicles within the cluster, i.e., on the cluster size N (Fig. 8) (Kerner and Klenov 2005, 2006a, b):

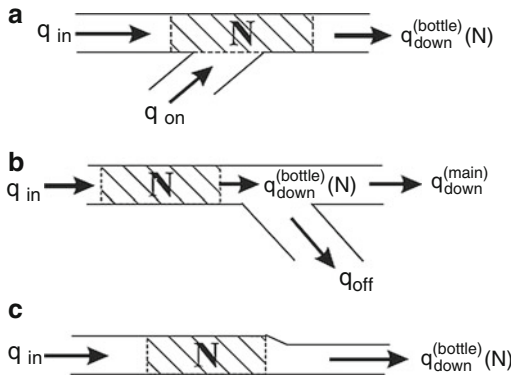
$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{bottle})}(N). \quad (39)$$

To determine $q_{\text{down}}^{(\text{bottle})}(N)$, we can use the flow rate dependence of the density (Fig. 4b). However, in the nucleation model for traffic breakdown (Kerner and Klenov 2005, 2006a, b), which we discuss below, we make an approximation in which we average the infinite synchronized flow states for each density to one average speed.



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 7 Schematic illustration of vehicle cluster transformation due to fluctuations in an initially

nonhomogeneous free flow at bottleneck (Adapted from Kerner and Klenov (2005, 2006a, b))



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 8 Qualitative explanation of vehicle cluster (dashed regions) at bottlenecks: (a) On-ramp bottleneck. (b) Off-ramp bottleneck. (c) Merge bottleneck

Thus, in this approximation, rather than Fig. 4, we use the related simplified nonmonotonous dependencies shown in Fig. 9.

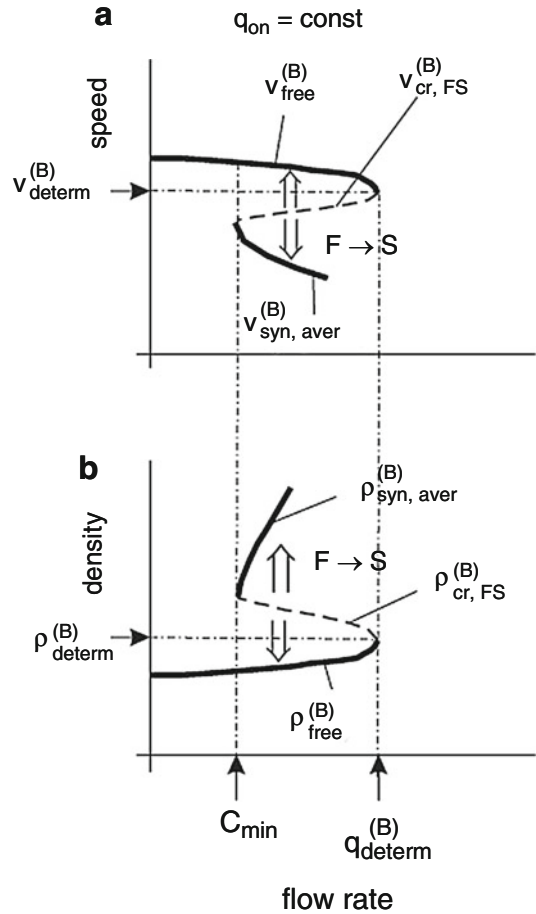
In the model, it is further assumed that the shape of the characteristic $q_{\text{down}}^{(\text{bottle})}(N)$ (Fig. 10) follows from the S-shaped density–flow characteristic shown in Fig. 9b. In this nucleation model, the characteristic $q_{\text{down}}^{(\text{bottle})}(N)$ has at least two different branches $q_{\text{down}}^{(\text{bottle})}(N)$ labeled $N^{(\text{determ})}$ and N_c in Fig. 10a. These branches are related to the vehicle number ranges, respectively, given by the conditions

$$0 \leq N \leq N_d \quad (40)$$

and

$$N_d < N \leq N_s. \quad (41)$$

The branches $N^{(\text{determ})}$ and N_c in Fig. 10a are associated with the density branches $\rho_{\text{free}}^{(B)}$ and $\rho_{\text{cr, FS}}^{(B)}$ of the S-shaped density–flow characteristic in Fig. 9b, respectively. The branch $N^{(\text{determ})}$ is associated with the case in which at a high enough flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ and the on-ramp flow rate $q_{\text{on}} > 0$ the deterministic cluster exists at the bottleneck under free flow conditions. The branch N_c is associated with the case in which the critical cluster whose growth leads to an $F \rightarrow S$ transition occurs at the bottleneck.

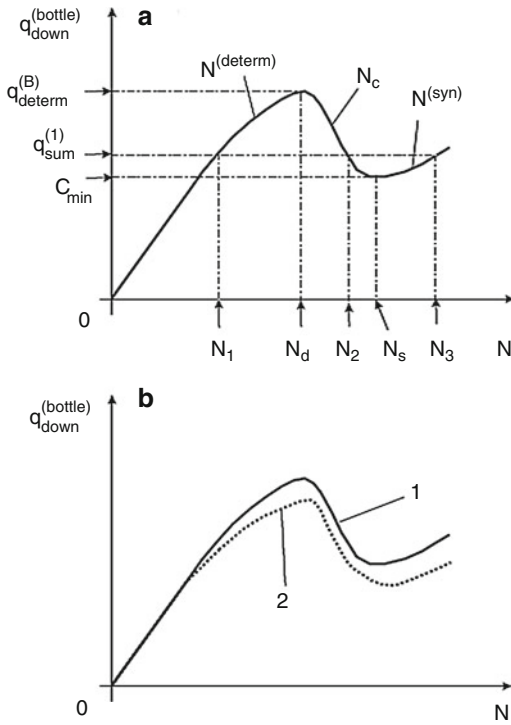


Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 9 Simplified qualitative shapes of speed (a) and density (b) dependencies on the flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ at a given value of the on-ramp inflow rate q_{on} associated with location of vehicle cluster at on-ramp bottleneck (Adapted from Kerner and Klenov (2005, 2006a, b))

In addition, from the S-shaped density–flow characteristic (Fig. 9b) can be seen that at

$$N > N_s, \quad (42)$$

there can be a third branch $N^{(\text{syn})}$ on the characteristic $q_{\text{down}}^{(\text{bottle})}(N)$ (Fig. 10a) associated with the branch $\rho_{\text{syn, aver}}^{(B)}$ for averaged synchronized flow states in Fig. 9b. In this case, $q_{\text{down}}^{(\text{bottle})}(N)$ Eq. 39 is a N-shaped flow–vehicle-number characteristic. It should be noted that the branch for synchronized flow $v_{\text{syn, aver}}^{(B)}$ in Fig. 9b and the associated branch



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 10 Qualitative dependencies of the

outflow rate $q_{\text{down}}^{(\text{bottle})}$ on the total vehicle number N within the cluster localized at the bottleneck (a), and possible dependencies of the N-shaped function $q_{\text{down}}^{(\text{bottle})}(N)$ on q_{on} for two different values $q_{\text{on}}^{(2)}$ (b); curve 1 for $q_{\text{on}} = q_{\text{on}}^{(1)}$, curve 2 for $q_{\text{on}} = q_{\text{on}}^{(2)} > q_{\text{on}}^{(1)}$. In (a), the flow rate $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ satisfies condition (8)

$N^{(\text{syn})}$ on the characteristic $q_{\text{down}}^{(\text{bottle})}(N)$ (Fig. 10a) follow from the microscopic three-phase theory (Kerner 2004; Kerner and Klenov 2003), rather than from the nucleation model of section “[Model of Traffic Breakdown at Highway Bottleneck](#).” The branch $N^{(\text{syn})}$, which can be drawn for the case only in which a localized SP (LSP) occurs as a result of traffic breakdown, is shown with the aim of a qualitative illustration of a possible traffic flow state after synchronized flow nucleation. Congested spatiotemporal traffic patterns, which cannot be described by the nucleation model of section “[Model of Traffic Breakdown at Highway Bottleneck](#),” are discussed in Kerner (2004, 2009b; entry ► “[Spatiotemporal Features of Traffic Congestion](#)”).

At the critical point $N = N_d$ at which the branches $N^{(\text{determ})}$ and N_c merge, the function $q_{\text{down}}^{(\text{bottle})}(N)$ has its maximum point. At the point $N = N_s$ at which the branches N_c and $N^{(\text{syn})}$ merge, the function $q_{\text{down}}^{(\text{bottle})}(N)$ has its minimum point.

In the case of the on-ramp bottleneck, quantitative characteristics of the N-shaped function $q_{\text{down}}^{(\text{bottle})}(N)$ (e.g., values N_d and N_s) can depend on the flow rate to the on-ramp q_{on} . This is because the on-ramp inflow and the flow rate q_{in} upstream of the bottleneck can make a considerable different influence on the cluster size and the outflow rate from the cluster $q_{\text{down}}^{(\text{bottle})}$. In particular, it can turn out that at the same N the larger q_{on} , the more difficult for vehicles to escape from the cluster, i.e., the less $q_{\text{down}}^{(\text{bottle})}$ is. This is confirmed by microscopic simulations (Kerner et al. 2002) and reflected in Fig. 10b in which it is assumed that the larger q_{on} , the less $q_{\text{down}}^{(\text{bottle})}$ and the larger N_d and N_s are. Thus, in a case of the on-ramp bottleneck instead of Eq. 39 we should use

$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}}), \quad (43)$$

where q_{on} is a parameter.

In accordance with the nucleation model (Fig. 10a), critical cluster occurrence describes an $F \rightarrow S$ transition at the bottleneck. There are two reasons for this statement: (i) A vehicle cluster is on average motionless, i.e., fixed at the bottleneck. This is also related to the definition of synchronized flow whose downstream front is usually fixed at the bottleneck, whereas the downstream front of a wide moving jam propagates through any bottleneck while maintaining the mean downstream jam velocity (Kerner 2004). (ii) The shape of the chosen function (39) in the nucleation model associated with this motionless cluster (Fig. 10a) follows from a Z-shaped speed–flow characteristic for traffic breakdown (Fig. 9a) found in the microscopic three-phase traffic theory (Kerner 2004; Kerner and Klenov 2003). In this theory, it has been shown that if these two requirements are satisfied, then rather than an $F \rightarrow J$ transition, an $F \rightarrow S$ transition occurs at the bottleneck, as found in empirical observations.

Steady States

Steady states of the vehicle number N within the cluster at a given flow rate q_{sum} are associated with solutions of the equation

$$q_{\text{sum}} = q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}}) \quad (44)$$

for the on-ramp bottleneck and

$$q_{\text{sum}} = q_{\text{down}}^{(\text{bottle})}(N) \quad (45)$$

for the off-ramp and merge bottlenecks. As can be seen from Fig. 10a, at a given flow rate $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ that satisfies condition (8) there can be at least two steady states: $N = N_1$ associated with the deterministic cluster and $N = N_2$ associated with the critical cluster. These steady states are the roots of Eq. 44, i.e., they are associated with the intersection points of the horizontal line $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ with the branches $N^{(\text{determ})}$ and N_c of the characteristic $q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}})$ (Fig. 10a), respectively. In addition, as mentioned above, if an LSP occurs as a result of an $F \rightarrow S$ transition, then there is a third root of Eq. 44, $N = N_3$, associated with the intersection point of the horizontal line $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ with the branch $N^{(\text{syn})}$ of the characteristic $q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}})$.

If the flow rate q_{sum} increases, then the critical vehicle number difference within the cluster

$$\Delta N_c = N_2 - N_1 \quad (46)$$

decreases. This critical vehicle number difference is associated with the difference between the number of vehicles within the critical cluster and the number of vehicles within the deterministic cluster at the bottleneck. The growth of the critical cluster leads to traffic breakdown at the bottleneck.

At the critical flow rate Eq. 5, we get $\Delta N_c = 0$: the steady states N_1 and N_2 merge into one point with the critical vehicle number $N = N_d$ at which

$$q_{\text{determ}}^{(B)} = q_{\text{down}}^{(\text{bottle})}(N_d, q_{\text{on}}). \quad (47)$$

At $q_{\text{sum}} \geq q_{\text{determ}}^{(B)}$ the deterministic traffic

breakdown should occur even if there were no random increase in the vehicle number within the deterministic cluster at the bottleneck.

If the flow rate q_{sum} decreases gradually, then the minimum highway capacity of free flow at the bottleneck

$$q_{\text{sum}} = C_{\text{min}} \quad (48)$$

is reached at which the steady states N_2 and N_3 merge into one steady state $N = N_s$ at which

$$C_{\text{min}} = q_{\text{down}}^{(\text{bottle})}(N_s, q_{\text{on}}). \quad (49)$$

Based on the nucleation model discussed above and on the master Eq. 33, we consider the dynamics of the vehicle cluster at a bottleneck (dashed regions in Fig. 8) (Kerner and Klenov 2005, 2006a, b). The probability $p(N, t)$ that N vehicles occur within the cluster at time t is given by the master Eq. 33 in which the variable x is the total vehicle number within the cluster N :

$$\begin{aligned} \frac{\partial p(N, t)}{\partial t} = & w_+(N-1)p(N-1, t) + w_-(N+1)p(N+1, t) \\ & - [w_+(N) + w_-(N)]p(N, t), \quad \text{at } N > 0. \end{aligned} \quad (50)$$

The master Eq. 50 describes probability p for the total vehicle number within the cluster, i.e., the cluster size N , which randomly changes due to fluctuations in a neighborhood of the size $N^{(\text{determ})}$ of the deterministic cluster (Fig. 7) (Kerner and Klenov 2005, 2006a, b). The size of the deterministic cluster $N^{(\text{determ})}$ does not depend on fluctuations. Random fluctuations cause either a decrease or an increase in the cluster size N in comparison with $N^{(\text{determ})}$. The fluctuations, which decrease the cluster size N , are associated with an increase in speed within the cluster. Therefore, these fluctuations can prevent traffic breakdown. In contrast, fluctuations, which increase the cluster size N , i.e., decrease the speed within the cluster, can cause traffic breakdown before the deterministic nucleus for traffic breakdown associated with condition (5) is reached.

In the master Eq. 50, w_+ and w_- are the attachment rate onto the cluster and detachment rate

from the cluster, respectively. The attachment rate onto the cluster w_+ does not depend on N :

$$w_+ = q_{\text{sum}}. \quad (51)$$

In contrast, the detachment rate from the cluster w_- is given by formula

$$w_- = q_{\text{down}}^{(\text{bottle})}. \quad (52)$$

Thus, the detachment rate from the cluster w_- depends on N in accordance with Eq. 39 for off-ramp and merge bottlenecks, or else in accordance with Eq. 43 for the case of an on-ramp bottleneck; in the latter case, w_- depends on N and q_{on} . When $N = 0$, the following boundary condition should be satisfied in Eq. 50

$$w_-(0) = 0, \quad (53)$$

i.e., no vehicles can leave the cluster. In this case,

$$\frac{\partial p(0,t)}{\partial t} = w_-(1)p(1,t) - w_+(0)p(0,t), \quad \text{at } N=0. \quad (54)$$

Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks

As follows from the analysis of the model (50)–(54) and formula (37) made in Kerner and Klenov (2005, 2006a, b), within the flow rate

range Eq. 8 the mean time delay of an $F \rightarrow S$ transition at the bottleneck is

$$T^{(B, \text{mean})} = 2\pi\tau_b \exp\{\Delta\Phi\}, \quad (55)$$

where a potential barrier for the $F \rightarrow S$ transition

$$\Delta\Phi = \Phi(N_2) - \Phi(N_1), \quad (56)$$

the potential $\Phi(N)$

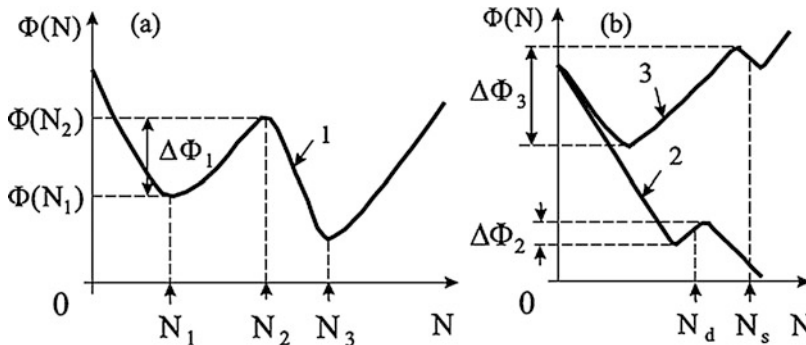
$$\Phi(N) = \begin{cases} \sum_{n=1}^N \ln \frac{w_-(n)}{w_+} & \text{at } N > 0, \\ 0 & \text{at } N = 0, \end{cases} \quad (57)$$

$$\tau_b = (w'_-(N_1)|w'_-(N_2)|)^{-\frac{1}{2}}, \quad (58)$$

$w'_-(N) = dw_-/dN$. Respectively, the nucleation rate for traffic breakdown is

$$G^{(B)} = \frac{1}{T^{(B, \text{mean})}} = \frac{1}{2\pi\tau_b} \exp\{-\Delta\Phi\}. \quad (59)$$

It can be seen from Eq. 55 that the mean time delay $T^{(B, \text{mean})}$ for traffic breakdown decreases exponentially with the increase in the potential barrier $\Delta\Phi$ Eq. 56. If in Fig. 11 the total flow rate increases from $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ to $q_{\text{sum}} = q_{\text{sum}}^{(2)}$ which is close to the critical flow rate Eq. 5 for deterministic traffic breakdown, then the potential barrier $\Delta\Phi$ Eq. 56 decreases from $\Delta\Phi_1$ to $\Delta\Phi_2$.



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 11 Qualitative shape of the potential $\Phi(N)$ Eq. 57 for different flow rates q_{sum} : curves 1, 2, and 3 are related to the corresponding flow rates $q_{\text{sum}}^{(1)}$, $q_{\text{sum}}^{(2)}$, and $q_{\text{sum}}^{(3)}$ satisfying the condition $q_{\text{sum}}^{(3)} < q_{\text{sum}}^{(1)} < q_{\text{sum}}^{(2)}$.

Characteristic values of the vehicle number within the cluster $N = N_i$, $i = 1, 2, 3$, $N = N_s$, and $N = N_d$ are discussed in section “Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory” (Fig. 10) (Adapted from Kerner and Klenov (2005, 2006a, b))

In contrast, if the flow rate q_{sum} decreases from $q_{\text{sum}} = q_{\text{sum}}^{(1)}$ to $q_{\text{sum}} = q_{\text{sum}}^{(3)}$, which is close to the minimum highway capacity of free flow at the bottleneck C_{min} Eq. 48, then the potential barrier $\Delta\Phi$ Eq. 56 increases from $\Delta\Phi_1$ to $\Delta\Phi_3$ (Fig. 11). At the point $q_{\text{sum}} = C_{\text{min}}$ Eq. 48, the potential barrier $\Delta\Phi(N)$ reaches the maximum value

$$\Delta\Phi = \Phi(N_s) - \Phi(N_{\text{min}}), \quad (60)$$

where $N_{\text{min}} = N_1$ at $q_{\text{sum}} = C_{\text{min}}$. As a result, the mean time delay $T^{(\text{B, mean})}$ Eq. 55 strongly increases as q_{sum} approaches the minimum highway capacity of free flow at the bottleneck C_{min} . Under condition (9), no traffic breakdown at the bottleneck is possible. This result does not depend on a random increase in the vehicle number within the cluster.

If in the vicinity of the critical vehicle number N_d the function $w_-(N)$ can be approximated by a parabolic function of N , then the following approximate formula can be derived from Eqs. 55–58:

$$T^{(\text{B, mean})} = \frac{\sqrt{2\pi}N_d}{q_{\text{determ}}^{(\text{B})}(\xi_d\Delta_c)^{1/2}} \left(\frac{1 + \Delta_c^{1/2}}{1 - \Delta_c^{1/2}} \right)^{2\sqrt{2/\xi_d}N_d} \times \exp\left(-\frac{4\sqrt{2}N_d\Delta_c^{1/2}}{\sqrt{\xi_d}}\right), \quad (61)$$

where

$$\xi_d = -(N^2 d^2 \ln w_- / dN^2)|_{N=N_d} \quad (62)$$

is a dimensionless value of the order of 1,

$$\Delta_c = \frac{q_{\text{determ}}^{(\text{B})} - q_{\text{sum}}}{q_{\text{determ}}^{(\text{B})}}, \quad (63)$$

i.e., Δ_c is the relative difference between the critical flow rate $q_{\text{determ}}^{(\text{B})}$ for the deterministic $F \rightarrow S$ transition and the flow rate q_{sum} Eq. 2. If in Eq. 61 $\Delta_c \ll 1$, then we get

$$T^{(\text{B, mean})} = \frac{\sqrt{2\pi}N_d}{q_{\text{determ}}^{(\text{B})}(\xi_d\Delta_c)^{1/2}} \exp\left(\frac{8N_d\Delta_c^{3-2}}{3\sqrt{2}\xi_d}\right). \quad (64)$$

Respectively, the nucleation rate $G^{(\text{B})} = 1/T^{(\text{B, mean})}$ for traffic breakdown at the bottleneck associated with Eq. 64 is

$$G^{(\text{B})} = \frac{q_{\text{determ}}^{(\text{B})}(\xi_d\Delta_c)^{1/2}}{\sqrt{2\pi}N_d} \exp\left(-\frac{8N_d\Delta_c^{3/2}}{3\sqrt{2}\xi_d}\right). \quad (65)$$

Note that $q_{\text{determ}}^{(\text{B})}$, N_d , and ξ_d for the on-ramp bottleneck depend on q_{on} ; in this case, $T^{(\text{B, mean})}$ Eq. 64 and $G^{(\text{B})}$ Eq. 65 are functions of q_{sum} and q_{on} .

As usual for each phase transition observed in many other metastable systems of natural science (Gardiner 1994), the nucleation rate for traffic breakdown Eq. 65 is an exponential function of Δ_c Eq. 63. For traffic flow, in accordance with Eqs. 65 and 63 the exponential growth of the nucleation rate with Δ_c Eq. 63 is very sensible to the value of the critical flow rate $q_{\text{determ}}^{(\text{B})}$ for the deterministic traffic breakdown. This emphasizes the important impact of the deterministic cluster, which occurs at the on-ramp bottleneck at $q_{\text{on}} > 0$, on the nucleation rate for traffic breakdown Eq. 65 at a given flow rate q_{sum} .

Probability of Traffic Breakdown

The mean time delay $T^{(\text{B, mean})}$ of traffic breakdown at the bottleneck calculated in section “Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks” enables us to find probability $P^{(\text{B})}$ that traffic breakdown occurs at the bottleneck during a time interval for observing traffic flow T_{ob} . Let us consider probability

$$P_C^{(\text{B})} = 1 - P^{(\text{B})} \quad (66)$$

that during a chosen time interval T_{ob} there is no $F \rightarrow S$ transition, i.e., free flow remains at the bottleneck. In accordance with a stochastic theory of dynamic metastable systems (Gardiner 1994), a dependence of the probability $P_C^{(\text{B})}$ on T_{ob} governs by the equation

$$\frac{dP_C^{(\text{B})}}{dT_{\text{ob}}} = -\frac{P_C^{(\text{B})}}{T^{(\text{B, mean})}}, \quad (67)$$

where $T^{(\text{B, mean})}$ is a constant at given values of q_{sum} and q_{on} . This equation derived with Kramer’s method (Gardiner 1994) is valid under condition that $T^{(\text{B, mean})}$ is long enough in comparison with a

relaxation time toward a metastable steady state with the deterministic cluster of the size $N^{(\text{determ})} = N_1$. This relaxation time is of the order of τ_b Eq. 58, i.e., Eq. 67 is valid, if

$$\tau_b \ll T^{(\text{B}, \text{mean})}. \quad (68)$$

Note that as follows from Eq. 55, condition (68) can be satisfied only if the potential barrier $\Delta\Phi$ in Eq. 55 is relatively high, i.e., $\Delta\Phi \gg 1$ Eq. 56.

In accordance with a general theory of phase transitions in metastable systems (Gardiner 1994), the solution of Eq. 67 is

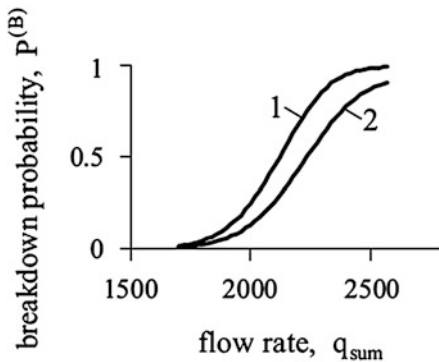
$$P_C^{(\text{B})}(T_{\text{ob}}) = P_0 \exp(-T_{\text{ob}}/T^{(\text{B}, \text{mean})}), \quad (69)$$

where P_0 is a constant. If the time interval T_{ob} is short enough, i.e., $T_{\text{ob}} \ll T^{(\text{B}, \text{mean})}$, but condition

$$\tau_b \ll T_{\text{ob}} \quad (70)$$

is also satisfied, then probability $P_C^{(\text{B})}$ should be close to one; therefore, $P_0 \approx 1$ in Eq. 69. Then under conditions (68), (70) probability $P_C^{(\text{B})}$ is given by the approximate formula

$$P_C^{(\text{B})}(T_{\text{ob}}) = \exp(-T_{\text{ob}}/T^{(\text{B}, \text{mean})}). \quad (71)$$



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 12 Probability of traffic breakdown $P^{(\text{B})}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}})$ calculated through formulas (61), (63), and (72) for different time intervals for observing traffic flow $T_{\text{ob}} = 30$ (curve 1) and 15 min (curve 2). $q_{\text{determ}}^{(\text{B})} = 2800$ vehicles/h, $N_d = 16$, $\xi_d = 2$

In this case, in accordance with Eq. 66 probability $P^{(\text{B})}$ of traffic breakdown at the bottleneck during the time interval T_{ob} is

$$P^{(\text{B})}(T_{\text{ob}}) = 1 - \exp(-T_{\text{ob}}/T^{(\text{B}, \text{mean})}). \quad (72)$$

Since $T^{(\text{B}, \text{mean})}$ in Eq. 72 is a function of q_{sum} and q_{on} , breakdown probability $P^{(\text{B})} = P^{(\text{B})}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}})$. Probability of traffic breakdown $P^{(\text{B})}$ for two different time intervals T_{ob} as functions of the flow rate q_{sum} calculated through formulas (61), (63), and (72) is shown in Fig. 12.

Stochastic Highway Capacity of Free Flow at Bottlenecks

Definition of Stochastic Highway Capacity

As well-known, highway capacity of free flow at a bottleneck is limited by traffic breakdown at the bottleneck (Brilon et al. 2005a, b; Brilon and Zurlinden 2004; Daganzo 1997; Eleftheriadou 2014, 2014, 1995; Gartner et al. 2001; Hall and Agyemang-Duah 1991; Hall et al. 1992; Greenshields et al. 1935; Lorenz and Eleftheriadou 2000; May 1990; Persaud et al. 1998). Traffic breakdown exhibits stochastic (probabilistic) nature associated with the nucleation nature of traffic breakdown (F $\rightarrow$ S transition) (Kerner 2004, 2017a) (section “Introduction”). Therefore, highway capacity is a stochastic value that definition should be consistent with the empirical nucleation nature of traffic breakdown. A critical discussion of the definition of highway capacity has already been made in Kerner (2017a, b). In this discussion, we have shown that the definition of highway capacity made in the three-phase theory (Kerner 2004, 2017a) is consistent with the empirical nucleation nature of traffic breakdown.

In the nucleation model of section “Model of Traffic Breakdown at Highway Bottleneck” (see Fig. 4), at any time instant there are the infinite number of the flow rates q_{sum} Eq. 10 in free flow at a highway bottleneck at which traffic breakdown can occur at the bottleneck. Therefore, these flow rates are highway capacities of free flow at the bottleneck.

This conclusion that is confirmed by empirical observations of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck (Kerner 2017a) leads to the following definition of stochastic highway capacity of free flow at a highway bottleneck made in the three-phase theory (Kerner 2004, 2017a):

At any time instant, there are the infinite number of highway capacities C of free flow at the bottleneck. The range of the infinite number of highway capacities is limited by the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$:

$$C_{\min} \leq C \leq C_{\max}, \quad (73)$$

where $C_{\min} < C_{\max}$. The existence of an infinite number of highway capacities at any time instant means that highway capacity is stochastic.

The physical sense of this capacity definition is as follows. Highway capacity is limited by traffic breakdown in an initial free flow at a highway bottleneck. In other words, any flow rate q_{sum} in free flow at the bottleneck at which traffic breakdown can occur is highway capacity. At any time instant, there are the infinite number of such highway capacities $C = q_{\text{sum}}$ at which traffic breakdown *can* occur. These capacities satisfy conditions (73).

Thus, in the three-phase theory any flow rate q_{sum} in free flow at a highway bottleneck that satisfies conditions

$$C_{\min} \leq q_{\text{sum}} \leq C_{\max} \quad (74)$$

is equal to one of the stochastic highway capacities of free flow at the bottleneck.

It should be noted that the model of the deterministic cluster in free flow at the bottleneck as well as the deterministic traffic breakdown that follows from this model considered in section “[Model of Traffic Breakdown Within Vehicle Cluster](#)” are a roughly simplification of a real complex spatiotemporal phenomena that determine features of a local disturbance in free flow at the bottleneck. Spatiotemporal phenomena within the local disturbance in free flow at the bottleneck that determine the nucleus emergence at the bottleneck have been considered in Sects.

5.12–5.14 of the book (Kerner 2017a). The spatiotemporal phenomena are characterized by a complex spatiotemporal competition between driver speed adaptation and driver over-acceleration. A consideration of these phenomena is out of scope of this entry.

Additionally, we should note that real fluctuations in free flow as well as the heterogeneity of real traffic flow that have been neglected in the model of the deterministic cluster can cause a considerable decrease in the maximal highway capacity $C_{\max}$ in comparison with the value $C_{\max}$ following from formula (6) (see the book (Kerner 2017a)). In other words, the value of the maximal highway capacity $C_{\max} = q_{\text{determ}}^{(B)}$ Eq. 6 can be considered a deterministic hypothetical limit value for the maximal highway capacity of free flow at the bottleneck that might be achieved if there were no fluctuations in free flow at the bottleneck. For this reason, when this deterministic hypothetical limit is discussed in presentations of results of numerical simulations (Figs. 13 and 14) considered below, to avoid confusions, rather than the maximal highway capacity $C_{\max}$ we use the notation $q_{\text{determ}}^{(B)}$.

Diagram of Traffic Breakdown

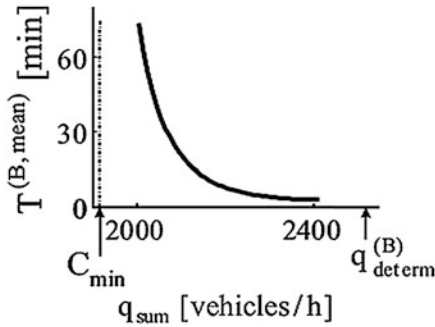
A link between the mean time delay $T^{(B, \text{mean})}$ of traffic breakdown and the probability $P^{(B)}$ of traffic breakdown discussed in section “[Probability of Traffic Breakdown](#)” is confirmed by numerical simulations presented in Figs. 13 and 14 (Kerner and Klenov 2005, 2006a, b). For the numerical simulations, we used an example of the function $w_{-}(N)$ Eq. 52

$$w_{-}(N) = N \left[\frac{a(q_{\text{on}})}{1 + (N/N_0(q_{\text{on}}))^4} + b(q_{\text{on}}) \right], \quad (75)$$

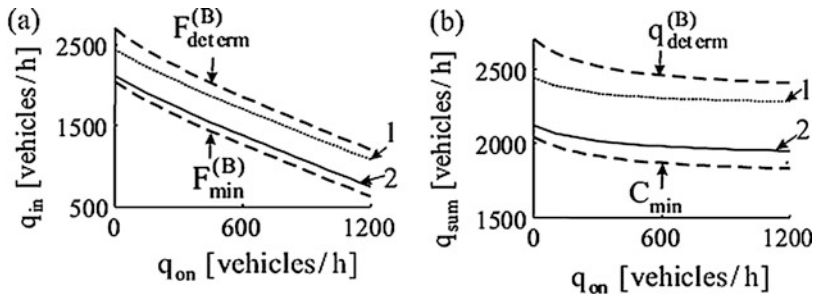
which exhibits qualitatively the same shape as that shown in Fig. 10. The analytical function (75) allows us to perform a numerical analysis of the mean time delay Eq. 55 and the associated nucleation rate Eq. 59 of traffic breakdown ($F \rightarrow S$ transition). For the analysis of Eqs. 55 and 59, only branches $N^{(\text{determ})}$ and N_c (Fig. 10) of the function

(75) associated with the deterministic and critical clusters within which $N \leq N_s$ are relevant. This is because the maximum possible value of $N = N_2$ in the potential barrier $\Delta\Phi$ Eq. 56 that determines the nucleation rate Eq. 59 is equal to N_s . As mentioned, the nucleation model cannot be applied for $q_{on} = 0$; therefore, in Fig. 14 points in the vicinity of $q_{on} = 0$ show only the tendency of the boundaries and the flow rates for the limiting case of small values q_{on} in which, however, the on-ramp inflow rate $q_{on} > 0$, specifically, it is assumed that the deterministic cluster still exists at the bottleneck.

A numerical study shows that the flow rate dependence of the mean time delay $T^{(B, \text{mean})}$ Eq. 55 (Fig. 13) exhibits qualitative features found



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 13 Mean time delay for traffic breakdown $T^{(B, \text{mean})}$ Eq. 55 as functions of the total flow rate q_{sum} for $q_{on} = 300$ vehicles/h



Mathematical Probabilistic Approaches to Traffic Breakdown, Fig. 14 Boundaries of constant values of breakdown probability $P^{(B)}$ in diagrams of traffic breakdown in the flow-flow planes $q_{in}(q_{on})$ (a) and $q_{\text{sum}}(q_{on})$ (b). $F^{(B)}_{\text{determ}}$ and $q^{(B)}_{\text{determ}}$ Eq. 5 are the critical boundary and critical flow rate for deterministic traffic breakdown, respectively.

in the microscopic three-phase theory (Kerner 2017a; Kerner and Klenov 2003; Kerner et al. 2002).

In the diagram of traffic breakdown at the bottleneck (Fig. 14a), there are boundaries $F^{(B)}_{S, \zeta}$ (curves 1 and 2) each of them is found from the condition that traffic breakdown occurs during the time T_{ob} with the probability $P^{(B)}$ that is equal to a given value ζ :

$$P^{(B)}(T_{ob}, q_{\text{sum}}, q_{on}) = \zeta, \quad \zeta = \text{const.} \quad (76)$$

We denote by $q^{(B)}_G$ the flow rate q_{sum} that is a solution of Eq. 76 at given T_{ob} and q_{on} :

$$q_{\text{sum}} = q^{(B)}_G. \quad (77)$$

A dependence of the flow rate $q^{(B)}_G(q_{on})$ on the on-ramp inflow rate q_{on} at given T_{ob} and ζ determines the boundary $F^{(B)}_{S, \zeta}$ (curves 1 and 2 in Fig. 14a).

The boundary $F^{(B)}_{S, \zeta}$ at $\zeta \approx 1$ (curve 1 in Fig. 14a) is qualitatively similar with the boundary $F^{(B)}_S$ found in numerical simulations of three-phase microscopic traffic flow models (Kerner 2004). In the diagram, there is also the boundary $F^{(B)}_{\text{min}}$ (curve $F^{(B)}_{\text{min}}$ in Fig. 14a) at which formula (48) is satisfied. In other words, at the boundary $F^{(B)}_{\text{min}}$ the flow rate q_{sum} is equal to the minimum highway capacity C_{min} . The shape of the boundary $F^{(B)}_{\text{min}}$ is determined by a dependence of the minimum highway capacity on the on-ramp inflow rate

C_{min} Eq. 48 is the minimum highway capacity; $F^{(B)}_{\text{min}}$ is the diagram boundary at which the minimum highway capacity is reached. Curves 1, 2 are related to different chosen values $P^{(B)}$ in Eq. 76: $\zeta = 0.986$ (curves 1) and $\zeta = 0.22$ (curves 2) at $T_{ob} = 15$ m

q_{on} . The shape of the boundary $F_{\text{determ}}^{(B)}$ is determined by the on-ramp inflow dependence of the critical flow rate $q_{\text{determ}}^{(B)}$ at which the hypothetical deterministic traffic breakdown occurs at the bottleneck.

Boundaries for traffic breakdown can also be calculated for another diagram of traffic breakdown whose co-ordinates are q_{sum} and q_{on} (Fig. 14b). Curves 1 and 2 in Fig. 14b show on-ramp inflow dependencies of the critical flow rate for traffic breakdown associated with two different values ζ of breakdown probability (76). In this diagram, the boundaries $q_{\text{determ}}^{(B)}$ and C_{min} are given by the on-ramp inflow dependencies of the critical flow rate for the deterministic traffic breakdown and the minimum highway capacity, respectively.

Conclusions

A mathematical probabilistic theory of traffic breakdown (Kerner 2000a, 2004; Kerner and Klenov 2005, 2006a, b) discussed in this entry explains the empirical nucleation nature of traffic breakdown that can be considered the empirical fundamental of transportation science (Kerner 2017a, b). The nucleation theory of traffic breakdown allows us to conclude that in accordance with measured traffic data traffic breakdown is a local phase transition from metastable free flow to synchronized flow ($F \rightarrow S$ transition). The $F \rightarrow S$ transition exhibits features of phase transitions observed in diverse complex metastable systems of natural science: (i) an $F \rightarrow S$ transition can be either spontaneous or induced; (ii) emergence and dissolution of synchronized flow is accompanied by a hysteresis effect; (iii) breakdown probability is an increasing function of control parameters (flow rates).

Future Directions

Whereas empirical *macroscopic* nucleation features of traffic breakdown have been found (see chapter 3 in the book (Kerner 2017a)), there are

almost no empirical investigations of *microscopic* empirical nucleation characteristics of traffic breakdown. In particular, behavior of speed fluctuations in real free flow at bottlenecks, appearance, and growth of critical fluctuations (critical vehicle clusters) leading to traffic breakdown have not been understood. In general, empirical and theoretical analyzes of various sources of traffic flow fluctuations and their behavior near traffic breakdown should be a very important direction in the field of traffic flow dynamics.

Although some *spatiotemporal* features of nuclei for traffic breakdown and synchronized flow are incorporated into the KKW CA traffic flow model (section “[Probabilistic Description of Traffic Breakdown with Three-Phase Cellular Automaton \(CA\) Traffic Flow Model](#)”), this is not made for the analytical study of traffic breakdown made in section “[Probabilistic Description of Traffic Breakdown Based on Master Equation](#).” In the latter case, to use the well-known probabilistic theory of phase transitions in metastable systems based on a master equation and one step processes (Gardiner 1994), possible various spatial shapes of nuclei have been ignored and assumed that the cluster dynamics can be described by the dynamic behavior of the vehicle number within the cluster only. Therefore, we believe that there is a great potential for further analytical probabilistic theories of traffic breakdown in which a spatiotemporal character of vehicle clusters at bottlenecks (see Sect. 5.12 and 5.13 in the book (Kerner 2017a)) is taken into account.

Acknowledgments We thank our partners for their support in the project “MEC-View – Object detection for automated driving based on Mobile Edge Computing,” funded by the German Federal Ministry of Economic Affairs and Energy.

Bibliography

- Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793–800
- Brilon W, Zurlinden H (2004) Kapazität von Straßen als Zufallsgröße. *Straßenverkehrstechnik* (4): 164
- Brilon W, Geistefeld J, Regler M (2005a) Reliability of freeway traffic flow: a stochastic concept of capacity.

- In: Mahmassani HS (ed) *Transportation and traffic theory*, Proceedings of the 16th international symposium on transportation and traffic theory. Elsevier, Amsterdam, pp 125–144
- Brilon W, Regler M, Geistefeld J (2005b) Zufallscharakter der Kapazität von Autobahnen und praktische Konsequenzen – Teil 1. *Straßenverkehrstechnik* (3): 136
- Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6:165–184
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
- Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
- Daganzo CF (1997) *Fundamentals of transportation and traffic operations*. Elsevier Science Inc, New York
- Elefteriadou L (2014) *An introduction to traffic flow theory*. Springer optimization and its applications, vol 84. Springer, Berlin
- Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Transp Res Rec* 1484:80–89
- Elefteriadou L, Kondyli A, Brilon W, Hall FL, Persaud B, Washburn S (2014) Enhancing ramp metering algorithms with the use of probability of breakdown models. *J Transp Eng* 140:04014003
- Gardiner CW (1994) *Handbook of stochastic methods for physics, chemistry, and the natural sciences*. Springer, Berlin
- Gartner NH, Messer CJ, Rathi A (eds) (2001) *Traffic flow theory. A state-of-the-art report*. Transportation Research Board, Washington, DC
- Gazis DC (2002) *Traffic theory*. Springer, Berlin
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7:499–505
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
- Greenshields BD, Bibbins JR, Channing WS, Miller HH (1935) A study of traffic capacity. *Highw Res Board Proc* 14:448–477
- Haight FA (1963) *Mathematical theories of traffic flow*. Academic, New York
- Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Transp Res Rec* 1320:91–98
- Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speedflow and flow-occupancy (or density) relationships on freeways. *Transp Res Rec* 1365:12–18
- Hausken K, Rehborn H (2015) Game-theoretic context and interpretation of Kerner's three-phase traffic theory. In: Hausken K, Zhuang J (eds) *Game theoretic analysis of congestion, safety and security*, Springer series in reliability engineering. Springer, Berlin, pp 113–141
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7:86–106
- Kerner BS (1998a) Theory of congested traffic flow. In: Rysgaard R (ed) *Proceedings of the 3rd symposium on highway capacity and level of service*, vol 2, Road Directorate, Ministry of Transport – Denmark, pp 621–642
- Kerner BS (1998b) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
- Kerner BS (1999a) Congested traffic flow: observations and theory. *Transp Res Rec* 1678:160–167
- Kerner BS (1999b) The physics of traffic. *Phys World* 12:25–30
- Kerner BS (2000a) Theory of breakdown phenomenon at highway bottlenecks. *Transp Res Rec* 1710:136–144
- Kerner BS (2000b) Experimental features of the emergence of moving jams in free traffic flow. *J Phys A Math Gen* 33:L221–L228
- Kerner BS (2001) Complexity of synchronized flow and related problems for basic assumptions of traffic flow theories. *Netw Spat Econ* 1:35–76
- Kerner BS (2002a) Synchronized flow as a new traffic phase and related problems for traffic flow modelling. *Math Comput Model* 35:481–508
- Kerner BS (2002b) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
- Kerner BS (2004) *The physics of traffic*. Springer, Berlin/New York
- Kerner BS (2009a) Traffic congestion, modelling approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9302–9355
- Kerner BS (2009b) Traffic congestion, spatiotemporal features of. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9355–9411
- Kerner BS (2009c) *Introduction to modern traffic flow theory and control*. Springer, Berlin/New York
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Phys A* 392:5261–5282
- Kerner BS (2015) Failure of classical traffic flow theories: a critical review. *Elektrotechn Informationstech* 132:417–433
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Phys A* 450:700–747
- Kerner BS (2017a) *Breakdown in traffic networks: fundamentals of transportation science*. Springer, Berlin/New York
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35:L31–L43
- Kerner BS, Klenov SL (2003) Microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68:036130
- Kerner BS, Klenov SL (2005) Probabilistic breakdown phenomenon at on-ramps bottlenecks in three-phase

- traffic theory. cond-mat/0502281, e-print in <http://arxiv.org/abs/cond-mat/0502281>
- Kerner BS, Klenov SL (2006a) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: congestion nucleation in spatially non-homogeneous traffic. *Phys A* 364:473–492
- Kerner BS, Klenov SL (2006b) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory. *Transp Res Rec* 1965:70–78
- Kerner BS, Klenov SL (2009) Phase transitions in traffic flow on multilane roads. *Phys Rev E* 80:056101
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35:9971–10013
- Kuhn TS (2012) The structure of scientific revolutions, 4th edn. The University of Chicago Press, Chicago/London
- Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2002) Probabilistic description of traffic breakdown. *Phys Rev E* 65:066125
- Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2004) Probabilistic description of traffic breakdown caused by on-ramp. E-print arXiv: cond-mat/0405163
- Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
- Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. *Trans Res C E-C018*:84–95
- Mahnke R, Kaupužs J (1999) Stochastic theory of freeway traffic. *Phys Rev E* 59:117–125
- Mahnke R, Pieret N (1997) Stochastic master-equation approach to aggregation in freeway traffic. *Phys Rev E* 56:2666–2671
- Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
- Mahnke R, Kaupužs J, Lubashevsky I (2009) Physics of stochastic processes. Wiley-VCH, Darmstadt
- May AD (1990) Traffic flow fundamentals. Prentice-Hall, Englewood Cliffs
- Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys (France) I* 2: 2221–2229
- Nagel K, Wagner P, Woesler R (2003) Still flowing: approaches to traffic flow and traffic jam Modeling. *Oper Res* 51:681–716
- Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Transp Res Rec* 1634:64–69
- Piccoli B, Tosin A (2009) In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer, Berlin, pp 9727–9749
- Rakha H, Wang W (2009) Procedure for calibrating Gipps car-following model. *Transp Res Rec* 2124:113–124
- Rakha H, Pasumarthy P, Adjerid S (2009) A simplified behavioral vehicle longitudinal motion model. *Transp Lett* 1:95–110
- Rehborn H, Klenov SL (2009) Traffic prediction of congested patterns. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer, Berlin, pp 9500–9536
- Rehborn H, Koller M (2014) A study of the influence of severe environmental conditions on common traffic congestion features. *J Adv Transp* 48:1107–1120
- Rehborn H, Palmer J (2008) ASDA/FOTO based on Kerner's three-phase traffic theory in North Rhine-Westphalia and its integration into vehicles. *IEEE Intelligent Veh Symp*:186–191. [http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=4607789&filter%3DAND\(p_IS_Number%3A4621124\)&pageNumber=3](http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=4607789&filter%3DAND(p_IS_Number%3A4621124)&pageNumber=3)
- Rehborn H, Klenov SL, Palmer J (2011a) An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. *Phys A* 390:4466–4485
- Rehborn H, Klenov SL, Palmer J (2011b) Common traffic congestion features studied in USA, UK, and Germany based on Kerner's three-phase traffic theory. *IEEE Intelligent Veh Symp IV*:19–24
- Rempe F, Franeck P, Fastenrath U, Bogenberger K (2016) Online freeway traffic estimation with real floating car data. In: Proceedings of 2016 I.E. 19th international conference on ITS, Rio de Janeiro, Brazil, November 1–4. pp 1838–1843
- Rempe F, Franeck P, Fastenrath U, Bogenberger K (2017) A phase-based smoothing method for accurate traffic speed estimation with floating car data. *Trans Res C* 85:644–663
- Schadschneider A, Chowdhury D, Nishinari K (2011) Stochastic transport in complex systems. Elsevier Science, New York
- Treiber M, Kesting A (2013) Traffic flow dynamics. Springer, Berlin
- Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
- Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
- Wolf DE (1999) Cellular automata for traffic simulations. *Phys A* 263:438–451

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory

Junfang Tian¹, Chenqiang Zhu¹ and Rui Jiang²

¹Institute of Systems Engineering, College of Management and Economics, Tianjin University, Tianjin, China

²MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing, China

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Three-Phase Traffic Theory

The Basic Models

The Three-Phase Traffic Flow Cellular

Automaton Models

Conclusion and Outlook

Bibliography

Glossary

CA models Cellular automata (CA) models are a class of microscopic traffic flow models. In the CA models, the time and space are discrete, and the evolution is described by update rules.

First-order phase transition In three-phase traffic theory, the $F \rightarrow S$ and $S \rightarrow J$ transitions are claimed to be first-order phase transition, in which the flow rate abrupt decreases.

Fundamental diagram Fundamental diagram describes the relationship between flow rate and density. In the empirical data, the flow rate and density are usually collected by loop detector and averaged over 1 min. In the simulation on a circular road, usually the global density vs. averaged flow rate is plotted.

Kerner's three-phase traffic theory In Kerner's three-phase theory, the congested flow has been further classified into synchronized flow (S) and

wide moving jam (J). Therefore, there are three phases in traffic flow, i.e., the free flow (F), synchronized flow, and wide moving jam. The usually observed traffic transitions from large empirical datasets across different countries are the transitions from free flow to synchronized flow ($F \rightarrow S$) and from synchronized flow to wide moving jam ($S \rightarrow J$). The direct $F \rightarrow J$ transition seldom occurs. Instead, the two step $F \rightarrow S \rightarrow J$ transition is observed: the $F \rightarrow S$ transition happens firstly, and then after some-time and at a different location, the $S \rightarrow J$ transition occurs.

Metastability In metastable density range, traffic flow can exhibit different state. A phase transition from one state to the other state might be induced by perturbation exceeding a threshold. The metastability is related to synchronized flow in three-phase models but is not related to synchronized flow in two-phase models, since synchronized flow is absent in two-phase models.

Nucleation nature of $F \rightarrow S$ transition The nucleation nature of $F \rightarrow S$ transition is the effect of the occurrence of a local perturbation in free traffic flow whose amplitude exceeds the critical amplitude. The subsequent growth of this perturbation leads to the $F \rightarrow S$ transition.

Nucleation nature of $S \rightarrow J$ transition The nucleation nature of $S \rightarrow J$ transition is the effect of the occurrence of a local perturbation in synchronized traffic flow whose amplitude exceeds the critical amplitude. The subsequent growth of this perturbation leads to the $S \rightarrow J$ transition.

Synchronized flow On the contrary, the downstream front of the synchronized flow is often fixed at a bottleneck. In synchronized flow, the average speed is lower than that in the free flow but greater than that in the wide moving jam, and the average flow is sometimes as high as in the free flow.

Two-phase traffic flow models In the two-phase traffic flow models, the traffic flow has been classified into free flow and congested flow. It is usually assumed that in the steady state, there

is a unique relationship between flow rate and density. When the traffic state deviates from the steady state, it has the tendency to return to the steady state. As a result, when the steady state is unstable, which usually occurs in the intermediate density range, homogeneous traffic flow cannot be maintained and thus transits into jams.

Wide moving jam Wide moving jam is a moving structure that maintains the mean speed of its downstream front, even as it propagates through other different traffic states of free flow and synchronized traffic flow or through highway bottlenecks. The downstream front of the jam is the region which separates the jam upstream and the free flow or the synchronized flow downstream. In wide moving jam, both the average speed and average flow is very low compared with that in the free flow.

Definition of the Subject and Its Importance

To understand the characteristics and mechanism of traffic flow evolution, many models have been proposed. In the two-phase models, the traffic flow has been classified into free flow (F) and congested flow. Kerner argued that the congested flow should be further classified into synchronized flow (S) and wide moving jam (J), and traffic breakdown at a highway bottleneck corresponds to a first-order $F \rightarrow S$ transition. He argued that two-phase traffic flow models are not able to explain the empirical nucleation nature of $F \rightarrow S$ transition at a highway bottleneck. Therefore, he proposed a three-phase traffic theory, which is claimed to be able to explain the empirical nucleation nature of traffic breakdown at freeway bottlenecks.

Cellular automata (CA) models are a class of microscopic traffic flow models. In the CA models, the time and space are discrete, and the evolution is described by update rules. The CA models have the advantage of flexible evolution rules and high computation efficiency. The development of three-phase CA models is an important contribution to the three-phase traffic theory.

Introduction

Traffic flow studies can be traced back to the 1930s (Greenshields et al. 1935). For nearly a century, many efforts have been devoted to the studies, which reveals many important and interesting traffic phenomena (Brackstone and McDonald 1999; Chowdhury et al. 2000; Helbing 2001; Kerner 2004, 2009, 2017; Treiber and Kesting 2013; Saifuzzaman and Zheng 2014; Zheng 2014; Jin et al. 2015), for example, (i) the relationship between flow and density, which is called fundamental diagram (Greenberg 1959; Pipes 1967). Many observations demonstrate that flow increases with the density until a maximum is reached. Then it begins to decrease with the density (Greenshields et al. 1935; May 1990; Brilon et al. 2005); (ii) capacity drop. When traffic congestion occurs, the throughput usually decreases by an order of about 10% (Hall and Agyemang-Duah 1991; Chung et al. 2007; Coifman and Kim 2011; Srivastava and Geroliminis 2013; Saberi and Mahmassani 2013); (iii) phantom jam. Jams can appear spontaneously in high-density traffic flow (Treiterer and Myers 1974; Sugiyama et al. 2008); (iv) traffic oscillations. In congested traffic, vehicles usually cannot maintain a steady speed, instead they alternate between slow-moving and fast-moving states (Daganzo and Daganzo 1997; Mauch and Cassidy 2002; Ahn and Cassidy 2007; Laval 2006; Laval and Daganzo 2006; Schönhof and Helbing 2007; Yeo and Skabardonis 2009; Laval and Leclercq 2010; Li and Ouyang 2011; Zheng et al. 2011; Li et al. 2012, 2014; Chen et al. 2012, 2014; Saifuzzaman et al. 2017); (v) wide scattering of flow-density data. While the flow-density data corresponds to a curve in the free flow, they are wide scattering in the congested flow (Kerner and Rehborn 1996, 1997; Kerner 2002; Treiber and Helbing 2003; Treiber et al. 2006).

To understand the characteristics and mechanism of traffic flow evolution, many models have been proposed, including the macroscopic ones (e.g., the seminal Lighthill-Whitham-Richards model (Lighthill and Whitham 1955; Richards 1956), the Payne model (Payne 1979), and the speed gradient model (Jiang et al. 2002)), and microscopic ones (e.g., the General Motor

(GM) models (Gazis et al. 1959, 1961; Chandler et al. 1958), the Gipps model (Gipps 1981), the optimal velocity (OV) model (Bando et al. 1995), the full velocity difference (FVD) model (Jiang et al. 2001), the intelligent driver (ID) model (Treiber et al. 2000)). According to Kerner (2013), these models can be roughly classified into two types: the two-phase model and the three-phase models. In the two-phase models, the traffic flow is classified into free flow and congested flow. It is usually assumed that in the steady state, there is a unique relationship between flow rate and density. When the traffic state deviates from the steady state, it has the tendency to return to the steady state. As a result, when the steady state is unstable, which usually occurs in the intermediate density range, homogeneous traffic flow cannot be maintained and thus transits into jams.

Recent experimental studies on car-following behaviors (Jiang et al. 2014, 2015, 2017) reveal that while fluctuations of the velocity and velocity difference of a leading car and a following car are small, the spacing between the two cars sometimes fluctuates in a wide range. More importantly, experiment demonstrates that the speed standard deviation grows in a concave way along the platoon, while it grows initially in a convex way in two-phase models, such as the GM model, the Gipps model, the OV model, the FVD model, the ID model, and so on. These findings demonstrate that the instability mechanism in two-phase models is debatable.

Kerner argues that the congested flow should be further classified into synchronized flow (S) and wide moving jam (J), and traffic breakdown at a highway bottleneck corresponds to a first-order $F \rightarrow S$ transition. He argues that two-phase traffic flow models are not able to explain the empirical nucleation nature of $F \rightarrow S$ transition at a highway bottleneck. Therefore, he proposes a three-phase traffic theory.

In three-phase theory, the $F \rightarrow S$ transition is supposed to be due to over-acceleration effect. The first three-phase model was proposed by Kerner and Klenov (2002). Later many different types of models were proposed. Comparing with the macroscopic models, the microscopic models focus on individual vehicles, whose movements are

assumed to be determined by the neighboring vehicles. The microscopic models based on differential equations are usually called car-following (CF) models, and models based on cellular automata (CA) approach are usually called CA models.

In the former type of models, the time and space are continuous, and the motion of cars is usually described by an ordinary differential equation of acceleration. In the CA models, the time, the space, and the speed are discrete, and the evolution is described by update rules. There are also difference equation models, where the time is discrete but the space and the speed are continuous.

The CA models have the advantage of flexible evolution rules and high computation efficiency. Therefore, ever since the proposal of the classical Nagel-Schreckenberg (NaSch) model (Nagel and Schreckenberg 1992), the CA approach develops very quickly, and hundreds of CA models have been presented. The previous several papers (Chowdhury et al. 2000; Maerivoet and Moor 2005; Knospe et al. 2004) have made good reviews of the development of CA models. Nevertheless, these papers were published more than 10 years ago and thus have not systematically reviewed the new development of three-phase CA models.

Motivated by the fact, we make a systematic review of the three-phase CA traffic flow models in this paper. We classify the models into different types based on the characteristics of their update rules and point out their merits and deficiencies. We hope the review could shed light on the further development of the three-phase CA models.

The paper is organized as follows. In section “[Three-Phase Traffic Theory](#),” three-phase traffic theory is very briefly reviewed. Section “[The Basic Models](#)” reviews several basic CA models. Section “[The Three-Phase Traffic Flow Cellular Automaton Models](#)” discusses the different types of three-phase CA models. Conclusion and outlook are given in section “[Conclusion and Outlook](#).”

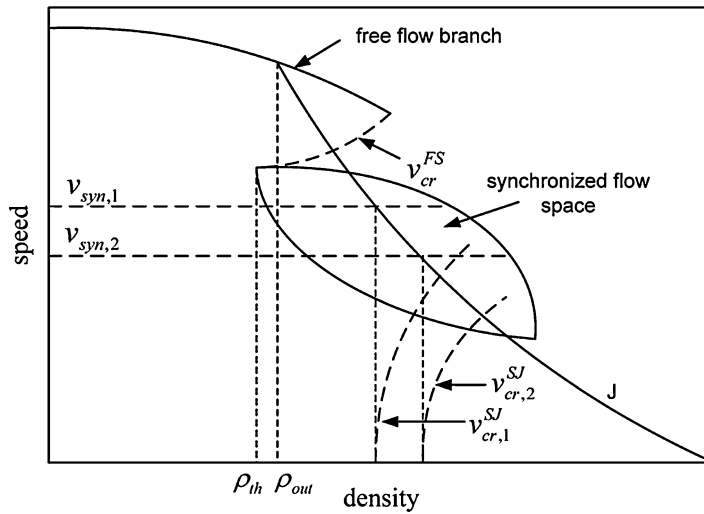
Three-Phase Traffic Theory

In Kerner’s three-phase theory (Kerner 1998, 1999a, b, c), the congested flow has been further classified into synchronized flow (S) and wide

moving jam (J). Therefore, there are three phases in traffic flow, i.e., the free flow (F), synchronized flow, and wide moving jam. Wide moving jam is a moving structure that maintains the mean speed of its downstream front, even as it propagates through other different traffic states of free flow and synchronized traffic flow or through highway bottlenecks. The downstream front of the jam is the region which separates the jam upstream and the free flow or the synchronized flow downstream. In wide moving jam, both the average speed and average flow are very low compared with that in the free flow. On the contrary, the downstream front of the synchronized flow is often fixed at a bottleneck. In synchronized flow, the average speed is lower than that in the free flow but greater than that in the wide moving jam, and the average flow is sometimes as high as in the free flow.

The phase transition in traffic flow is the process from one phase to another. The usually observed traffic transitions from large empirical datasets across different countries are the transitions from free flow to synchronized flow ($F \rightarrow S$), i.e., traffic congestion occurs, and from synchronized flow to wide moving jam ($S \rightarrow J$), i.e., traffic jams happen. The direct $F \rightarrow J$ transition seldom occurs, unless the formation of synchronized flow is strongly hindered due to a non-homogeneity, such as at a traffic split on a highway (Kerner 2000). Instead, the two step $F \rightarrow S \rightarrow J$ transition is observed: the $F \rightarrow S$ transition happens firstly, and then after sometime and at a different location, the $S \rightarrow J$ transition occurs (see Fig. 1), in which the $F \rightarrow S \rightarrow J$ transition shows the double Z-characteristic on the speed-density plane.

As introduced in section “History of the Emergence of Kerner’s Three-Phase Traffic Theory” in



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 1

The illustration of $F \rightarrow S \rightarrow J$ transition on the speed-density diagram. “J” is related to “line J” for the same 2Z-characteristic shown in the flow-density plane, whereas in the speed-density plane the line J becomes a curve. ρ_{th} is the minimum density where the $F \rightarrow S$ transition could happen. ρ_{out} is the density of the free outflow of the wide moving jams. The free flow branch, the synchronized flow space, and the $F \rightarrow S$ transition curve v_{cr}^{FS} form the first Z structure. $v_{cr}^{FS} = v_{free} - \Delta v_{cr}^{FS}$ and Δv_{cr}^{FS} are the amplitude of the critical disturbance for the $F \rightarrow S$ transition. v_{free} is the vehicle’s speed in free traffic flow. When a local perturbation occurs with the amplitude

$\Delta v_{dis}^{FS} > \Delta v_{cr}^{FS}$, i.e., its amplitude is greater than the critical amplitude Δv_{cr}^{FS} , this perturbation will grow and lead to an $F \rightarrow S$ transition. The synchronized flow space, the $S \rightarrow J$ transition curve v_{cr}^{SJ} , and the speed zero within wide moving jams form the second Z structure. $v_{cr}^{SJ} = v_{syn} - \Delta v_{cr}^{SJ}$ and Δv_{cr}^{SJ} are the amplitude of the critical disturbance for the $S \rightarrow J$ transition. v_{syn} is the vehicle’s speed in synchronized traffic flow. Since there are infinitely many different speeds in synchronized flow, there are also infinitely many threshold $S \rightarrow J$ transition curves v_{cr}^{SJ} for these different synchronized flow speeds. Only two of them, i.e., $v_{cr,1}^{SJ}$ and $v_{cr,2}^{SJ}$, are shown for two speeds $v_{syn,1}$ and $v_{syn,2}$

the entry ► “Traffic Prediction of Congested Patterns” by Rehborn et al. in this Encyclopedia, the reason to propose the three-phase theory is that “for more than 25 years the engineers have tried to implement ideas of classical traffic flow theory as well as traffic control methods based on these theories in real traffic installations: none of the methods and theories are in agreement with real field traffic data.”

Kerner claimed that real traffic breakdown at freeway bottleneck is a $F \rightarrow S$ phase transition that exhibits a nucleation nature. It is claimed that the classical traffic flow theories fail to while the three-phase traffic theory can explain the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ phase transition) at freeway bottlenecks.

The Basic Models

This section reviews three basic CA models. In the CA models, the road is classified into cells. Each cell can be empty or occupied by at most one vehicle (except in multi-value CA models, in which each cell can hold more than one vehicle, see, e.g., Nishinari and Takahashi (2000)). Time is also discretized and each time step corresponds to 1 s throughout this paper. In this review paper, we focus on the car-following behavior on a single circular lane. The lane changing behavior is another important topic and is left for another review paper. In most CA models, the movements of vehicles are usually updated in parallel.

The NaSch Model

The update rules in the NaSch model are as follows:

Step 1. Acceleration:

$$\tilde{v} = \min(v + a, v_{\max}).$$

Step 2. Deceleration:

$$\tilde{v} = \min(\tilde{v}, d).$$

Step 3. Randomization:

$$v' = \begin{cases} \min(\tilde{v} - b, 0), & \text{if } r < p, \\ \tilde{v}, & \text{otherwise.} \end{cases}$$

Step 4. Vehicle movement:

$$x' = x + v'.$$

Here x (x') and v (v') denote speeds and positions at the current and next time step. x_l is the position of the leading vehicle. $d = x_l - x - L_{\text{veh}}$ is the space gap between the two vehicles and L_{veh} is the vehicle length. $v_{\max}$ is the maximum speed of vehicles. a and b are the acceleration and randomization deceleration, respectively, which are usually set to be 1. p is the randomization probability.

Figure 2 shows the fundamental diagram of the NaSch model on a circular road with length 1000 L_{cell} . The parameters are shown in Table 1. Two different traffic states are classified. When the density is smaller than the critical density ρ_c , the traffic is in free flow (see Fig. 3a). When $\rho > \rho_c$, the jams appear spontaneously (see Fig. 3b).

The Velocity Effect Model

In the NaSch model, the preceding car is regarded as motionless, which might also move. To model this phenomenon, the anticipation effect is usually introduced. A velocity effect (VE) model was proposed by Li et al. (2001), which is a typical model that considers the anticipation effect. Note that the anticipation effect was also introduced earlier in the brake light model (see section “Brake Light Models”). The update rules are different from NaSch model only in the deceleration step,

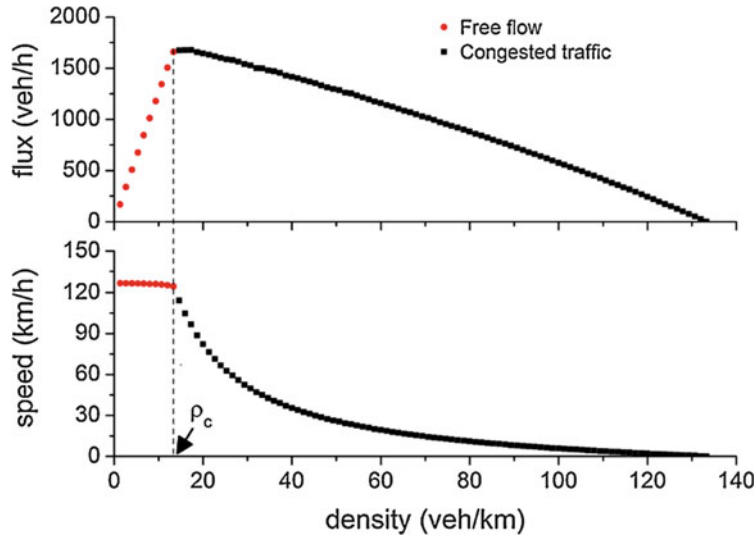
Step 2. Deceleration:

$$\tilde{v} = \min(\tilde{v}, d + v_l^{\text{anti}}).$$

where $v_l^{\text{anti}} = \min(v_{\max} - b, v_l, \max(0, d_l - b))$ is the anticipation speed of the preceding car. Figure 4 compares the fundamental diagrams of the NaSch model and the VE model. One can see that with the same parameters, the maximum flow rate in VE model is larger due to anticipation effect.

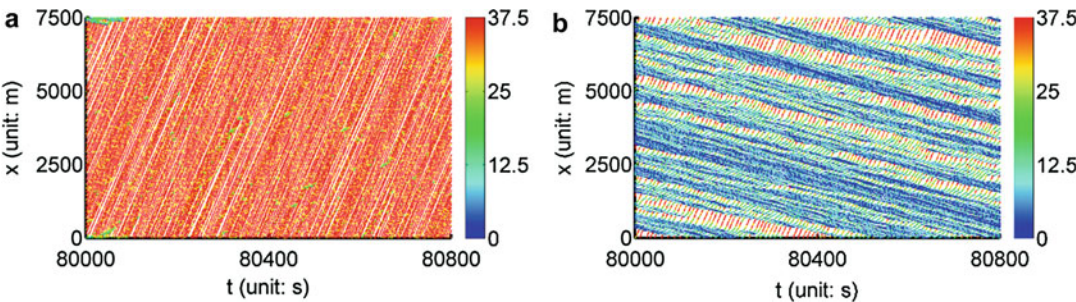
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 2

The fundamental diagram of the NaSch model on a circular road. This figure describes the averaged flow as a function of the global density (number of vehicles divided by the circumference). The averaged flow is the product of the global density and the average speed of all vehicles on the road



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 1 Parameter values of the NaSch model. (Taken from Nagel and Schreckenberg (1992))

Parameters	L_{cell}	L_{veh}	v_{max}	a	b	p
Units	m	L_{cell}	L_{cell}/s	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	—
Value	7.5	1	5	1	1	0.3



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 3 The spatiotemporal diagrams of NaSch model on the circular road.

(a) $\rho = 13\text{veh/km}$; (b) $\rho = 53\text{veh/km}$. The color bar indicates speed (m/s)

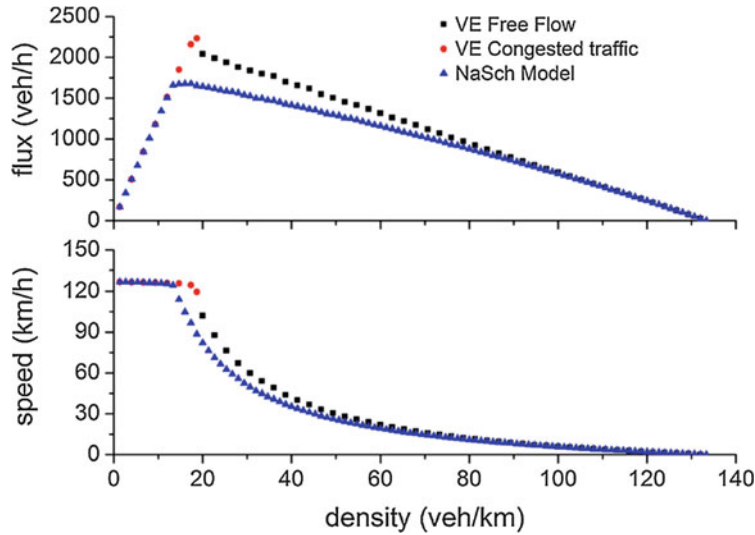
The Slow-to-Start Models

In the NaSch model, the transition from free flow to congested flow is not of first-order. To depict the first-order transition in traffic flow, the slow-to-start rules are introduced (Takayasu and Takayasu 1993; Benjamin et al. 1996; Fukui and Ishibashi 1997; Schadschneider and

Schreckenberg 1997; Barlovic et al. 1998). Here we review the velocity-dependent-randomization (VDR) rule (Barlovic et al. 1998), which is the simplest one. In the VDR model, the only difference from the NaSch model is that the randomization is not a constant; it depends on the velocity as follows:

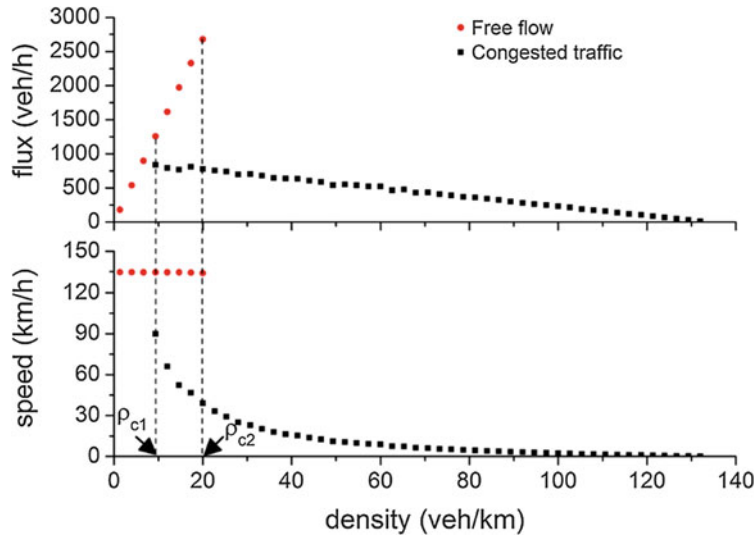
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 4

A comparison of the fundamental diagrams between the VE model and the NaSch model on a circular road



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 5

The fundamental diagram of the VDR model on a circular road. The parameters are shown in Table 2



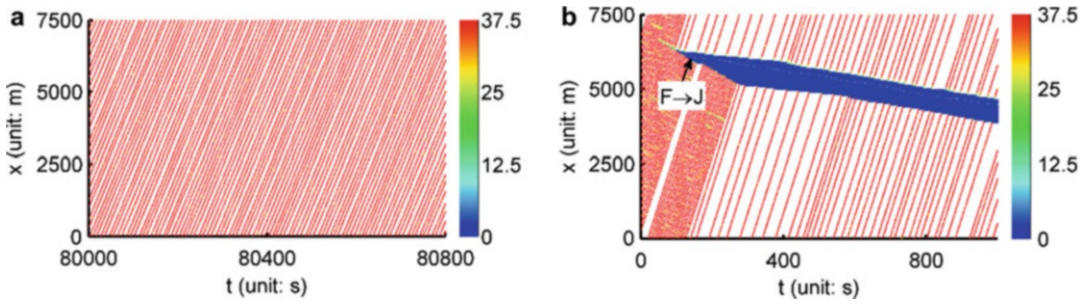
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 2 Parameter values of the VDR model. (Taken from Barlovic et al. (1998))

Parameters	L_{cell}	L_{veh}	v_{max}	a	b	p_0	p
Units	m	L_{cell}	L_{cell}/s	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	–	–
Value	7.5	1	5	1	1	0.75	1/64

$$p(v) = \begin{cases} p_0, & \text{if } v = 0, \\ p, & \text{otherwise.} \end{cases}$$

Figure 5 shows the fundamental diagram of the VDR model, in which there are two critical

densities. Below ρ_{c1} , the traffic is in free flow; above ρ_{c2} , the traffic is in jam. When the density is in the range $\rho_{c1} < \rho < \rho_{c2}$, the traffic flow is metastable. When starting from homogeneous traffic, free flow is maintained (see Fig. 6a).



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 6 The spatiotemporal diagrams of VDR model on the circular road. (a)

$\rho = 15 \text{ veh/km}$, start from homogeneous initial condition; (b) $\rho = 20 \text{ veh/km}$

When starting from megajam, the traffic is in jam. Moreover, the first-order transition from free flow to jam could appear spontaneously (see Fig. 6b).

The Three-Phase Traffic Flow Cellular Automaton Models

Models Considering Speed Adaptation Effect Explicitly

Kerner-Klenov-Wolf Model

The Kerner-Klenov-Wolf (KKW, Kerner et al. 2002) model is the one of the first CA models (the other is the comfortable driving model in section “Brake Light Models”) within the framework of three-phase traffic theory, in which the 2D region of steady state and the speed adaptation effect are explicitly considered. The update rules consist of the deterministic and stochastic rules.

Step 1. Dynamical part of all KKW models:

$$\tilde{v} = \max(0, \min(v_{\max}, v_s, v_c)).$$

where,

$$v_s = d.$$

$$v_c = \begin{cases} v + a & \text{for } x_l - x > G(v), \\ v + a \text{sign}(\Delta v) & \text{for } x_l - x \leq G(v). \end{cases}$$

Step 2. Stochastic part of all KKW models:

$$v' = \max(0, \min(\tilde{v} + a\eta, v + a, v_{\max}, v_s)).$$

where,

$$\eta = \begin{cases} -1, & \text{if } r < p_b, \\ 1, & \text{if } p_b \leq r < p_b + p_a, \\ 0, & \text{otherwise.} \end{cases}$$

$$p_b = \begin{cases} p_0, & \text{if } v = 0, \\ p, & \text{if } v > 0. \end{cases}$$

$$p_a = \begin{cases} p_{a1}, & \text{if } v < v_p, \\ p_{a2}, & \text{if } v \geq v_p. \end{cases}$$

Step 3. Vehicle movement:

$$x' = x + v'$$

Here $\Delta v = v_l - v$ is the speed difference and v_l is the speed of the leading vehicle. v_s is the safe speed that must not be exceeded to avoid collisions. v_c describes the rule of “speed change,” which means the acceleration behavior depends on whether the leading vehicle is within a synchronized space gap. The synchronized space gap can be defined as linear form: $G(v) = L_{\text{veh}} + kv$, or nonlinear form, $G(v) = L_{\text{veh}} + v + \beta v/(2a)$, in which k and β are the positive parameters. The sign function is defined as

$$\text{sign}(\Delta v) = \begin{cases} -1 & \text{if } \Delta v < 0, \\ 0 & \text{if } \Delta v = 0, \\ 1 & \text{if } \Delta v > 0. \end{cases}$$

The steady state corresponds to a 2D region bounded by $q = \rho v_s = (1 - \rho L_{\text{veh}})$ (U: related to

safe speed), $q = \rho v_{\max}$ (F, related to maximum speed), and $q = (1 - \rho L_{\text{veh}})/k$ (L, related to the linear form of synchronized space gap) (see Fig. 7). The rule $v_c = v + a \text{sign}(\Delta v)$ corresponds to the speed adaptation effect, which occurs when the traffic state is in the 2D region.

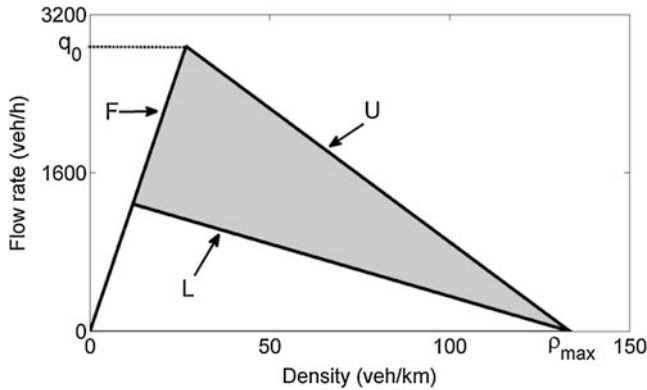
Figure 8 shows simulation results of the fundamental diagram of the KKW model on a circular road with length $L = 7500$ m. Each cell corresponds to $L_{\text{cell}} = 0.5$ m. The parameter values are shown in Table 3. One can see that three density ranges are classified by two critical densities. When the density $\rho < \rho_{c1}$, the traffic flow is in free flow state (see Fig. 9a). When the density is larger than ρ_{c1} , traffic flow is a coexistence of free flow and jams if initially starting from a megajam (see Fig. 9b). When the density $\rho_{c1} < \rho < \rho_{c2}$, the

traffic flow will be in synchronized flow if starting from a homogeneous configuration (see Fig. 9c). When $\rho > \rho_{c2}$, synchronized flow cannot be maintained even if starting from homogeneous configuration. Jams will emerge spontaneously (see Fig. 9d). We also notice that in the simulation results, the flow rate in synchronized flow changes non-monotonically, which might be related to the choice of parameters.

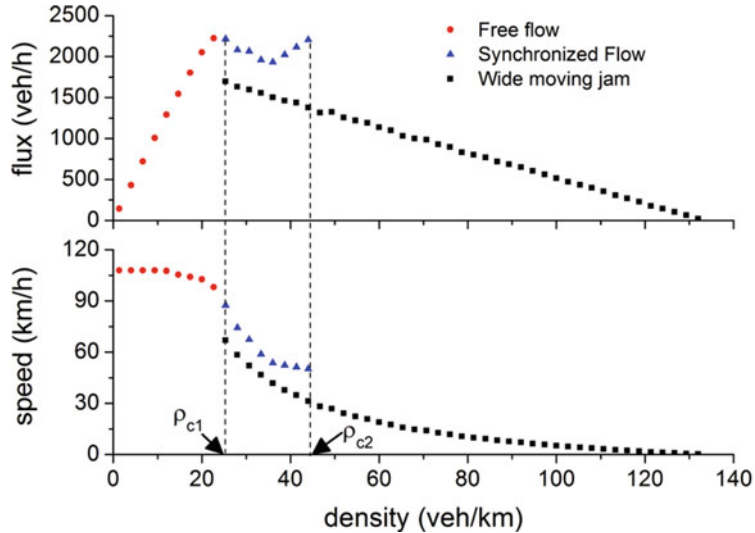
KKS Model

Kerner (2013) argued that in real traffic, there are at least two mechanisms for the over-acceleration effect. One is related to lane changing, and the other is related through driver acceleration without lane changing. Inspired by this fact, they developed the KKS model, which is named

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 7 The 2D region of KKW model

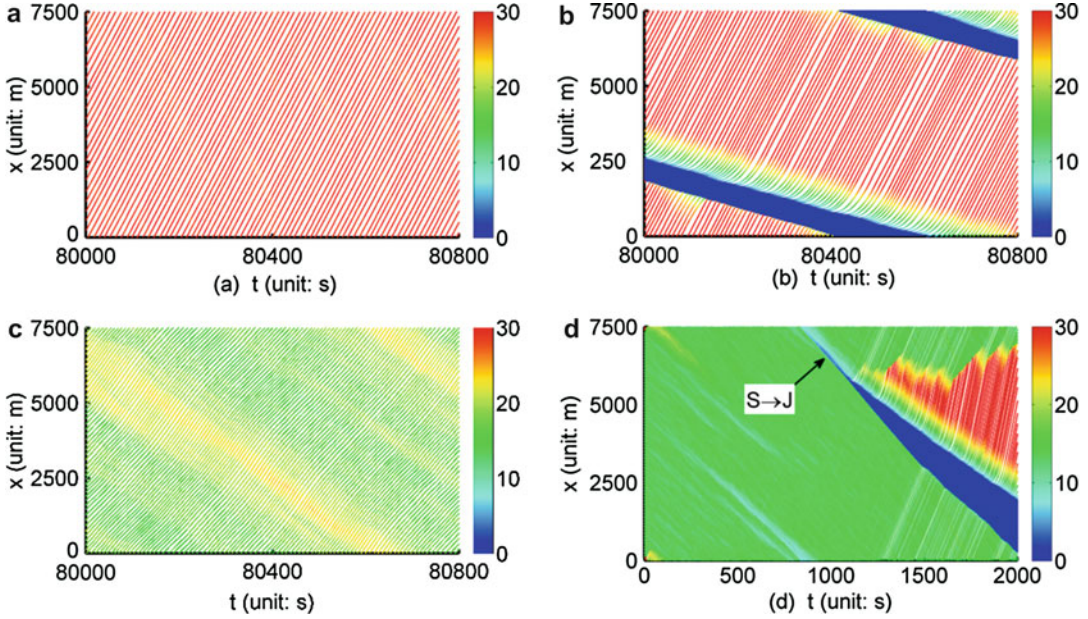


Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 8 The fundamental diagram of the KKW model on a circular road. The parameters are shown in Table 3



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 3 Parameter values of the KKW model. (Taken from Kerner et al. (2002))

Parameters	L_{cell}	L_{veh}	v_{max}	a	k	p	p_0	p_{a1}	p_{a2}	v_p
Units	m	L_{cell}	L_{cell}/s	$L_{\text{cell}}/\text{s}^2$	—	—	—	—	—	L_{cell}/s
Value	0.5	15	60	1	2.55	0.04	0.425	0.2	0.052	28

**Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 9** The spatiotemporal diagrams of KKW model on the circular road.

(a) $\rho = 15 \text{ veh/km}$; (b) $\rho = 31 \text{ veh/km}$; (c) $\rho = 31 \text{ veh/km}$; (d) $\rho = 47 \text{ veh/km}$

because the model transforms into the Kerner-Klenov-Schreckenberg (KKS, Kerner et al. 2011) model when the parameter $p_a = 0$, and it can be reduced to the KKW model for a single-lane road. The version of KKS model on a single lane is as follows:

Step 3. Over-acceleration through random acceleration within the synchronization space gap:

if $\Delta v \geq 0$ and $r < p_a$ then $\tilde{v} = \min(\tilde{v} + a, v_{\text{max}})$.

Step 4. Acceleration:

$$\tilde{v} = \min(v + a, v_{\text{max}}).$$

Step 1. Comparison of vehicle gap with the synchronization gap and velocity update:

if $d \leq G(v)$ then follow steps 2 and 3 and skip step 4;

if $d > G(v)$ then skip steps 2 and 3 and follow step 4.

Step 2. Speed adaptation within the synchronization space gap:

Step 5. Deceleration:

$$\tilde{v} = \min(\tilde{v}, d).$$

Step 6. Randomization:

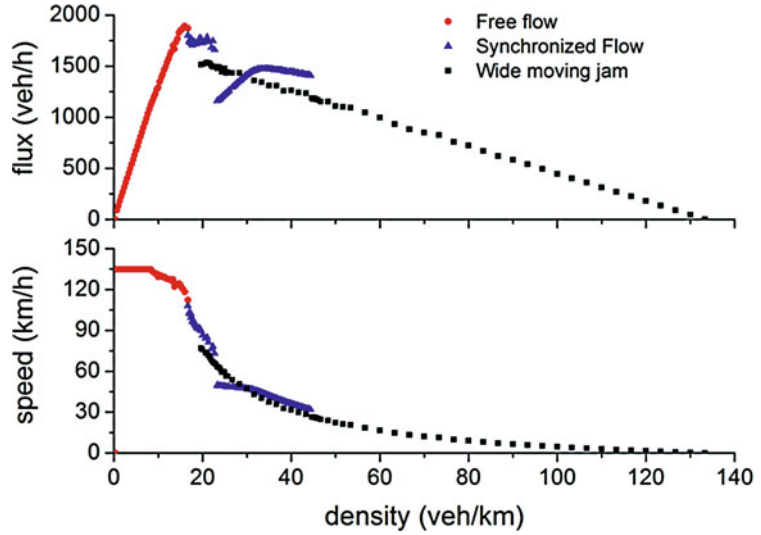
$$v' = \begin{cases} \max(\tilde{v} - b, 0), & \text{if } p_a < r < p_a + p_n, \\ \tilde{v}, & \text{otherwise.} \end{cases}$$

Step 7. Motion

$$\tilde{v} = v + \text{sign}(\Delta v).$$

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 10

The fundamental diagram of the KKS model on a circular road. The parameters are shown in Table 4



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 4 Parameter values of the KKS model. (Taken from Kerner (2013))

Parameters	L_{cell}	L_{veh}	v_{max}	a	b	k_1	k_2	p_3
Units	m	L_{cell}	L_{cell}/s	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	—	—	—
Value	1.5	5	25	1	1	3	2	0.01
Parameters	$p_0^{(2)}$	$p_2^{(2)}$	$p_{a,1}$	$p_{a,2}$	v_p	v_{syn}	Δv_{syn}	
Units	—	—	—	—	L_{cell}/s	L_{cell}/s	L_{cell}/s	
Value	0.5	0.35	0.07	0.08	8	14	3	

$$x' = x + v'.$$

Within the synchronization gap ($d_n \leq G(v_n)$), the driver adapts the vehicle speed to that of the preceding vehicle. Moreover, the vehicle might also over-accelerate through random acceleration with probability p_a , which is assumed to be an increasing function of speed:

$$p_a(v) = p_{a,1} + p_{a,2} \max(0, \min(1, (v - v_{\text{syn}})/\Delta v_{\text{syn}})).$$

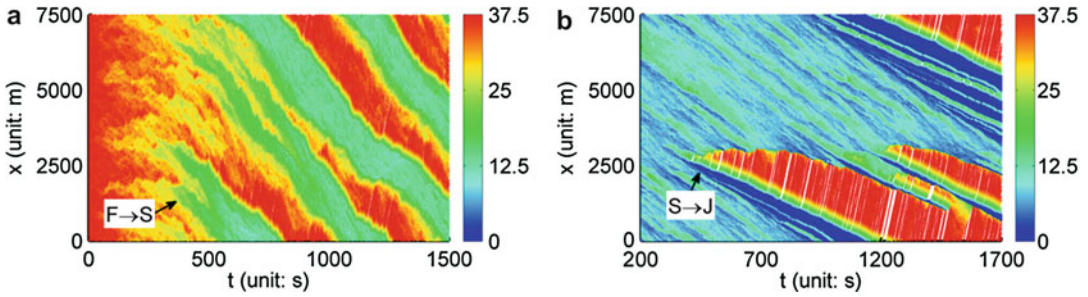
where $p_{a,1}$, $p_{a,2}$, v_{syn} , and Δv_{syn} are constants. The probability p_n is chosen as follows:

$$p_n = \begin{cases} p_2^{(2)}, & \text{if } \tilde{v} > v \text{ and } v > 0 \text{ and } v \leq v', \\ 0, & \text{if } \tilde{v} > v \text{ and } v > 0 \text{ and } v > v', \\ p_0^{(2)}, & \text{if } \tilde{v} > v \text{ and } v = 0, \\ p_3, & \text{if } \tilde{v} \leq v. \end{cases}$$

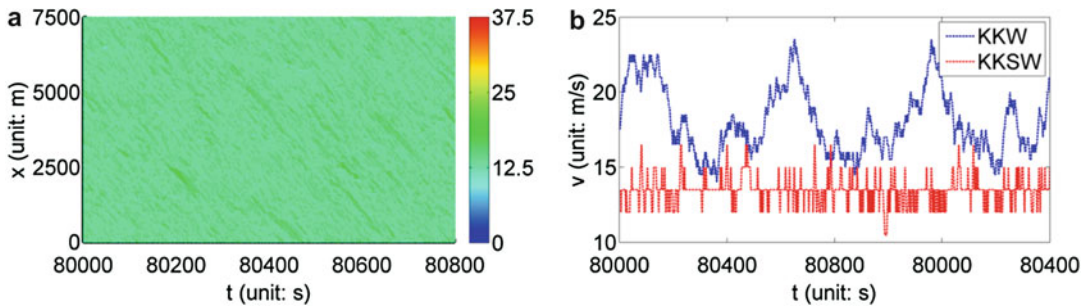
where v' is the speed in the last time step, and $\tilde{v}$ is determined in step 5. The synchronization gap $G(v) = k(v)v$, and the parameter

$$k(v) = \begin{cases} k_1, & \text{if } v > v_p, \\ k_2, & \text{otherwise.} \end{cases}$$

Figure 10 shows simulation results of the fundamental diagram of the KKS model on a circular road with length $L = 7500$ m. One can see that the fundamental diagram is different from KKW model, in which the synchronized traffic flow branch includes two subbranches. In the first one, where the densities are small, traffic flow and synchronized flow coexist (see Fig. 11a). In the second subbranch, only synchronized flow exists (see Fig. 12a). Surprisingly, flow rate abruptly drops between the two subbranches.



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 11 (a) $F \rightarrow S$ transition in the model of KKS, $\rho = 21 \text{ veh/km}$; (b). $S \rightarrow J$ transition in the model of KKS, $\rho = 49 \text{ veh/km}$



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 12 (a) The spatiotemporal diagram of the KKS model on a circular road. The

density $\rho = 27 \text{ veh/km}$; (b) speed of a vehicle in the KKS model ($\rho = 31 \text{ veh/km}$) and the KKS model ($\rho = 27 \text{ veh/km}$)

The $F \rightarrow S$ and $S \rightarrow J$ transitions can be reproduced by KKS model (see Fig. 11a, b). In the $F \rightarrow S$ transition, many local transitions gradually emerge. This is different from that in other models shown later, in which the transition abruptly occurs at one location and then spreads.

Finally, we compare the time series of the speed of a vehicle between KKS model and KKS model in Fig. 12b. One can see that the synchronized flow in KKS model is much more homogenous.

Generalized NaSch Models

Models of this class (such as Gao et al. 2007, 2009; Zhao et al. 2009; Jia et al. 2011; Tian et al. 2012a, b, 2015, 2016) are all based on the following generalized NaSch rules:

Step 1. Acceleration/deceleration:

$$\tilde{v} = \min(v + \hat{a}\hat{d}, \hat{v}).$$

Step 2. Randomization with probability $\hat{p}$:

$$v' = \begin{cases} \min(\tilde{v} - \hat{v}_r, 0), & \text{if } r < \hat{p}, \\ \tilde{v}, & \text{otherwise.} \end{cases}$$

Step 3. Vehicle movement:

$$x' = x + v'.$$

In different models, the formulations of $\hat{a}$, $\hat{d}$, $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ are different.

The Model of Gao et al.

In the model of Gao et al. (2007), the formulations of $\hat{a}$, $\hat{d}$, $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ are as follows:

- (i) $\hat{a} = a$
- (ii) $\hat{d} = d$
- (iii) $\hat{v} = v_{\max}$
- (iv)
$$\hat{v}_r = \begin{cases} a & \text{when } t_{st} \geq t_c, \\ b_- & \text{if } v < v_l \\ b_0 & \text{if } v = v_l \\ b_+ & \text{if } v > v_l \end{cases} \text{ otherwise.}$$
- (v)
$$\hat{p} = \begin{cases} p_0 & \text{when } t_{st} \geq t_c, \\ p_d & \text{when } t_{st} < t_c. \end{cases}$$

Here t_{st} denotes the time that a vehicle stops,

$$t'_{st} = \begin{cases} t_{st} + 1 & \text{if } v = 0, \\ 0 & \text{if } v > 0. \end{cases}$$

where t_c is a slow-to-start parameter; b_- , b_0 , and b_+ are the randomization decelerations corresponding to different velocity difference conditions; and $b_+ \geq a \geq b_-$ should be satisfied to simulate the synchronized flow. Note that the

speed adaptation effect is implicitly considered in the formulation of $\hat{v}_r$.

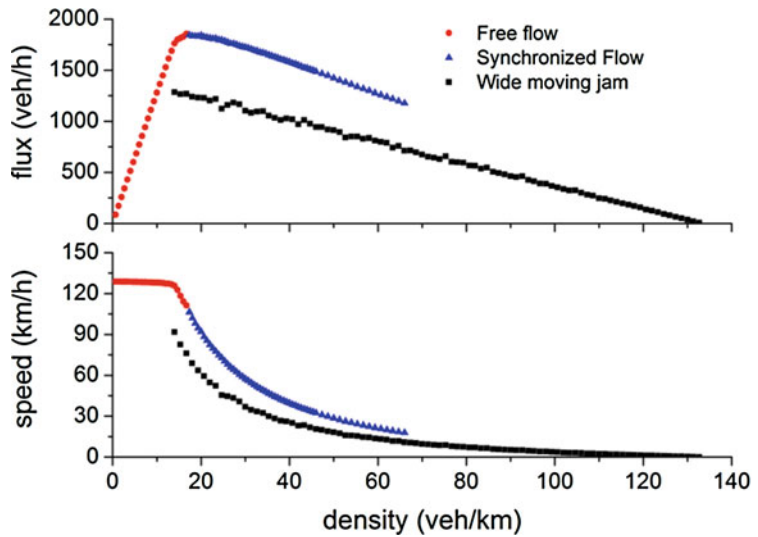
Figure 13 shows simulation results of the fundamental diagram of the Gao model, which is similar to that of KKW model, except that the synchronized flow branch monotonically decreases with the density. This model could reproduce the synchronized flow, the jam, and the $S \rightarrow J$ transition (see Fig. 14). However, it fails to reproduce the $F \rightarrow S$ transition.

To overcome this deficiency, Gao et al. (2009) later proposed an improved model, where $\hat{v}_r$ and $\hat{p}$ are changed into:

- (iv)
$$\hat{v}_r = \begin{cases} a & \text{when } t_{st} \geq t_c, \\ b_- & \text{if } v < v_l \\ b_0 & \text{if } v = v_l \\ b_+ & \text{if } v > v_l \end{cases} \begin{cases} \text{when } t_{st} < t_c \\ \text{and } d < D \\ \text{otherwise.} \end{cases}$$
- (v)
$$\hat{p} = \begin{cases} p_0 & \text{when } t_{st} \geq t_c, \\ p_d & \text{when } t_{st} < t_c \text{ and } d \leq D, \\ p_s & \text{when } t_{st} < t_c \text{ and } d > D. \end{cases}$$

Here D is an effective operating distance in velocity adaptation mechanism. Figure 15 shows

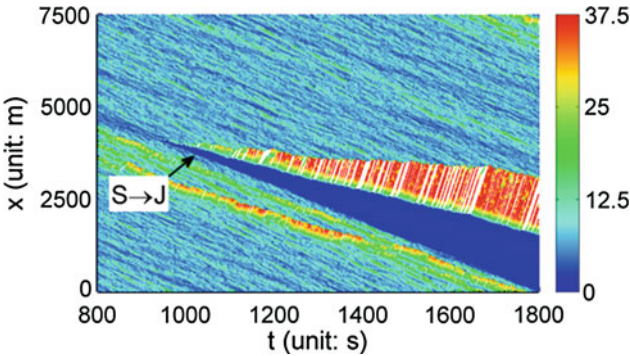
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 13 The fundamental diagram of Gao model on the circular road with length $L = 7500$ m. The parameters are shown in Table 5



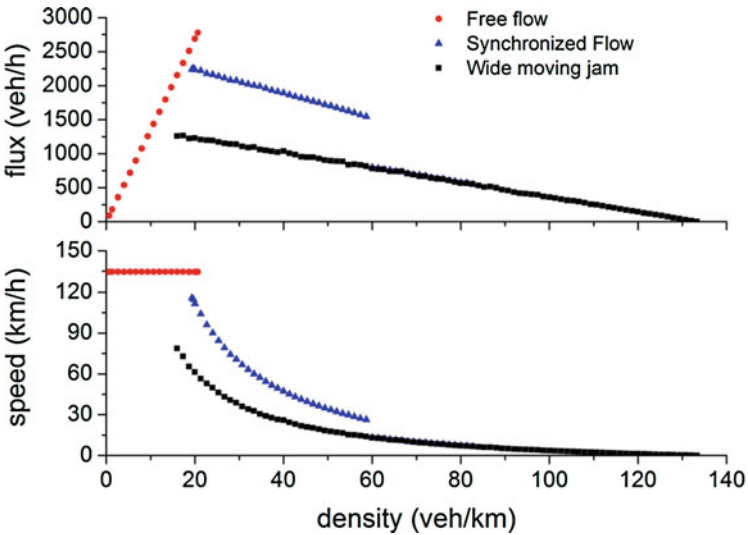
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 5 Parameter values of Gao model. (Taken from Gao et al. (2007))

Parameters	L_{cell}	L_{veh}	v_{max}	t_c	p_d	p_0	a	b	b_0	b_+
Units	m	L_{cell}	L_{cell}/s	s	—	—	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$
Value	1.5	5	25	7	0.3	0.6	2	1	2	5

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 14 $S \rightarrow J$ transition in the Gao model, $\rho = 69\text{veh/km}$



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 15 The fundamental diagram of improved Gao model on the circular road with length $L = 7500$ m. The parameters are shown in Table 6



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 6 Parameter values of improved Gao model. (Taken from Gao et al. (2007))

Parameters	L_{cell}	L_{veh}	v_{max}	t_c	p_d	p_0	p_s
Units	m	L_{cell}	L_{cell}/s	s	—	—	—
Value	1.5	5	25	6	0.18	0.5	0.08
Parameters	a	b	b_s	b_0	b_+	D	
Units	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	L_{cell}	
Value	2	1	1	2	5	23	

simulation results of the fundamental diagram of the improved Gao model. This model could reproduce the synchronized flow and the $F \rightarrow S$ and the $S \rightarrow J$ transitions (see Fig. 16).

We would like to mention that in the synchronized flow branch, the traffic flow is a coexistence of free flow and synchronized flow (see Fig. 16a). With the increase of density, the free flow region shrinks and finally vanishes (see Fig. 17a). We also point out that in the synchronized flow in the improved Gao model, the speed fluctuation of vehicles is too strong, which seems unrealistic (see Fig. 17b).

Two-State Model and the Improved One

In the two-state model (TS model), the formulations of $\hat{a}$, $\hat{d}$, $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ are as follows:

(i)

$$\hat{a} = a$$

(ii)

$$\hat{d} = d + \max(v_{\text{anti}} - g_{\text{safety}}, 0)$$

(iii)

$$\hat{v} = v_{\text{max}}$$

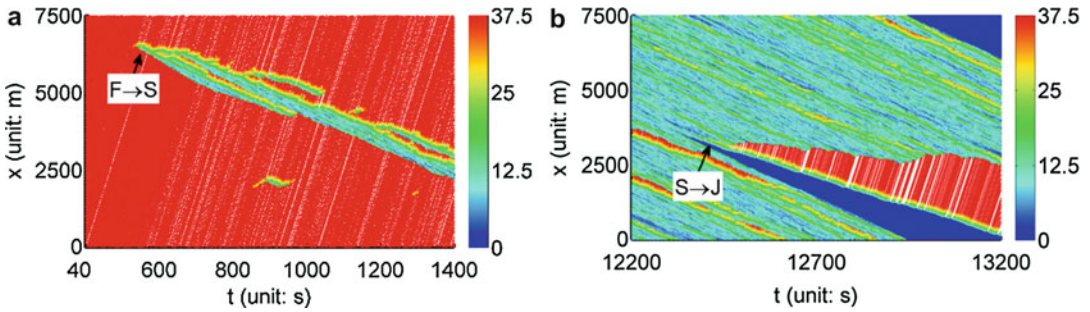
(iv)

$$\hat{v}_r = \begin{cases} a & \text{if } v \leq d_{\text{anti}}/T, \\ b_{\text{defense}} & \text{otherwise.} \end{cases}$$

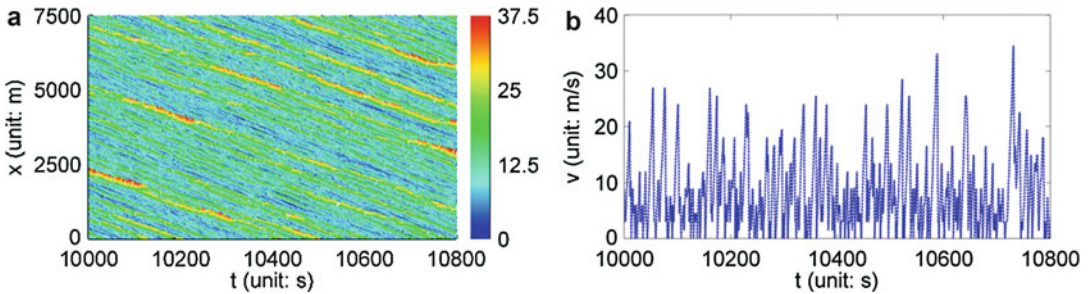
(v)

$$\hat{p} = \begin{cases} p_b & \text{if } v = 0 \text{ and } t_{st} > t_c, \\ p_c & \text{else if } 0 \leq v \leq d_{\text{anti}}/T, \\ p_a & \text{otherwise.} \end{cases}$$

Here $v_{\text{anti}} = \min(d_l, v_l + a, v_{\text{max}})$ is the anticipated speed of the leading vehicle, in which d_l is



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 16 The simulation results of the spatiotemporal diagram. (a) $\rho = 20 \text{ veh/km}$; (b) $\rho = 59 \text{ veh/km}$

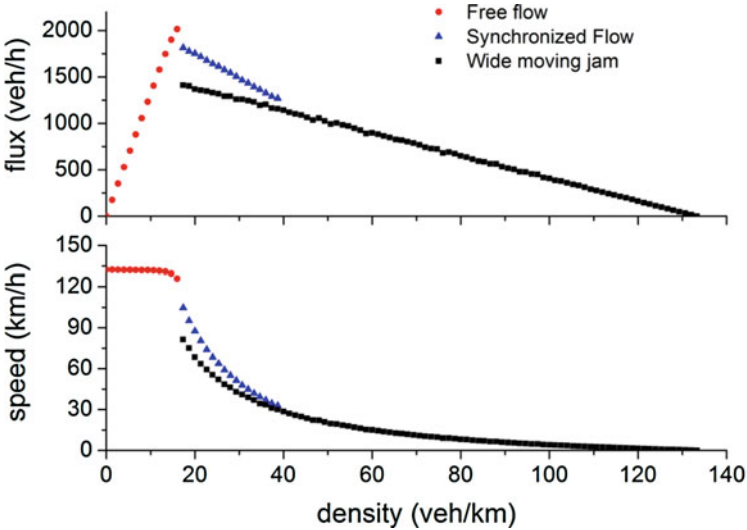


Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 17 The simulation results of spatiotemporal diagram. (a) $\rho = 55 \text{ veh/km}$; (b) the time series of velocity of a single vehicle

the space gap of the leading vehicle. The effectiveness of the anticipation is controlled by g_{safety} . Accidents are avoided only if $\hat{v}_r \geq g_{\text{safety}}$.

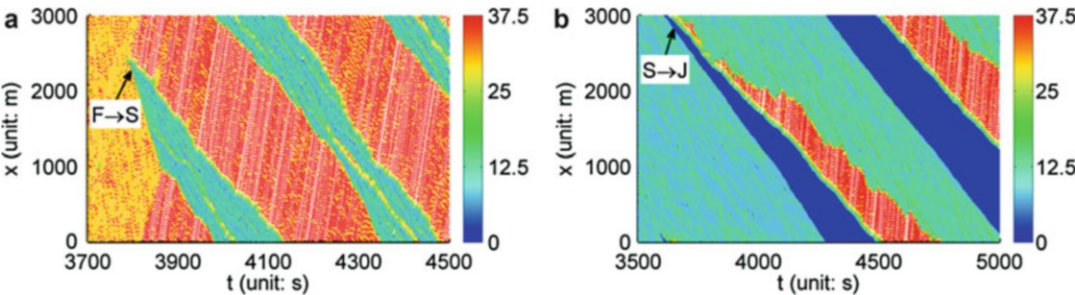
The TS model has considered two driver states: a defensive one and a normal one. The defensive state is activated if $v > d_{\text{anti}}/T$. Otherwise the drivers are assumed to be in the normal driving state.

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 18 The fundamental diagram of the TS model on a circular road with length $L = 3000$ m. The parameters are shown in Table 7



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 7 Parameter values of TS model. (Taken from Tian et al. (2015))

Parameters	L_{cell}	L_{veh}	v_{max}	T	p_a	p_b	p_c	a	b_{defense}	g_{safety}	t_c
Units	m	L_{cell}	L_{cell}/s	s	—	—	—	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	L_{cell}	s
Value	7.5	1	5	1.8	0.95	0.55	0.1	1	1	2	8



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 19 The spatiotemporal diagrams of TS model with the density $\rho = 25\text{veh/km}$ (a) and (b) $\rho = 67\text{veh/km}$

For more details, one may refer to Tian et al. (2016). To make the two-state model more realistic, the safe speed v_{safe} and the logistic function for the randomization probability p_{defense} are introduced into the two-state model. The two new variables are defined as follows:

$$v_{\text{safe}} = \left[-b_{\text{max}} + \sqrt{b_{\text{max}}^2 + v_l^2 + 2b_{\text{max}}d_n} \right].$$

$$p_{\text{defense}} = p_c + \frac{p_a}{1 + e^{\alpha(v_c - v_n)}}.$$

where α and v_c are the steepness and midpoint of the logistic function, respectively. v_l is the speed of the preceding car; b_{max} is the maximum deceleration of the car. The round function $[x]$ returns the integer nearest to x . Since safe speed is introduced, the model is named as the two-state model

with safe speed (TSS model, Tian et al. 2016). In the TSS model, the formulations of $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ change into:

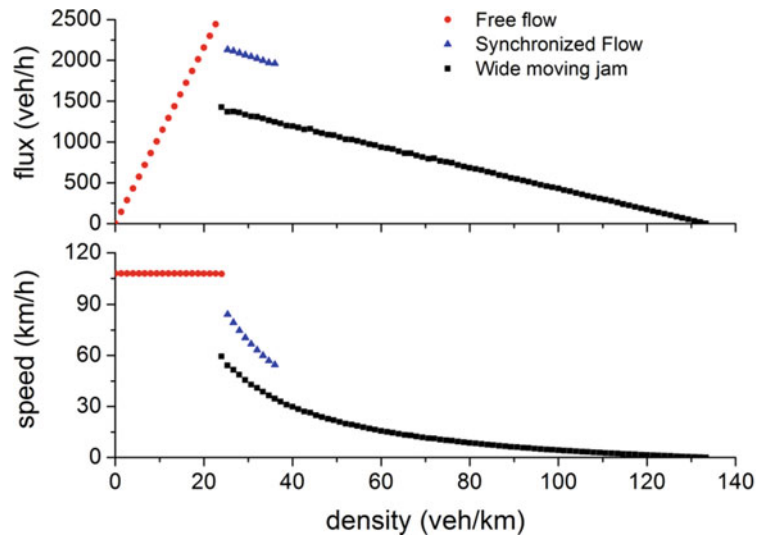
$$\hat{v} = \min(v_{\text{max}}, v_{\text{safe}}).$$

$$\hat{v}_r = \begin{cases} a & \text{if } v < b_{\text{defense}} + [d_{\text{anti}}/T], \\ b_{\text{defense}} & \text{otherwise.} \end{cases}$$

$$\hat{p} = \begin{cases} p_b & \text{if } v = 0, \\ p_c & \text{else if } 0 < v \leq d_{\text{anti}}/T, \\ p_{\text{defense}} & \text{otherwise.} \end{cases}$$

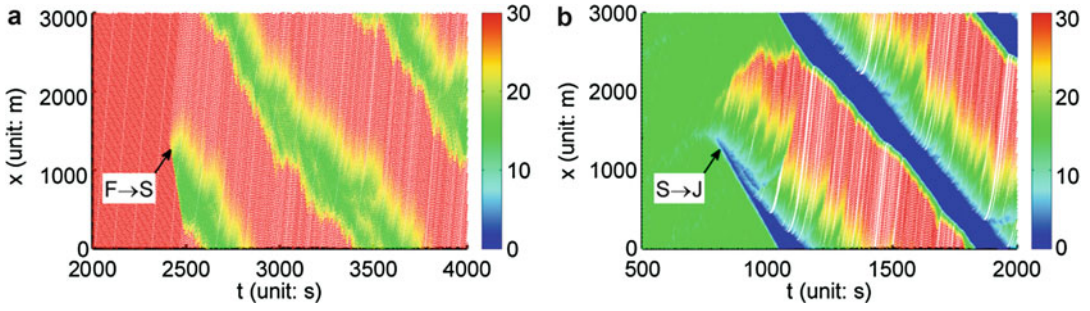
Here $[x]$ returns the maximum integer no greater than x . Figure 20 shows the simulation results of the fundamental diagram of the improved two-state model. One can see that the model can well reproduce the three phases of

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 20 The fundamental diagram of the improved two-state model on a circular road with length 3000 m. The parameters are shown in Table 8

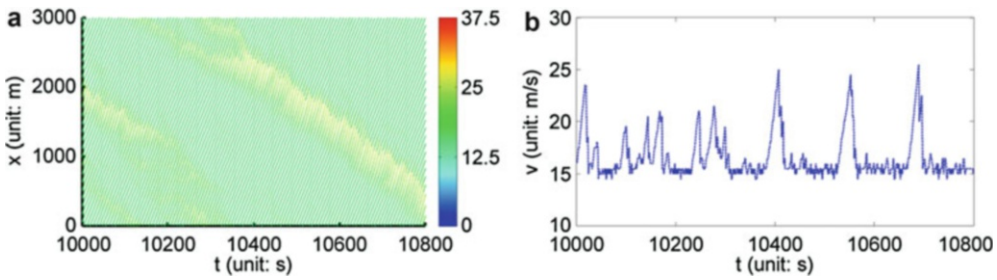


Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 8 Parameter values of TSS model. (Taken from Tian et al. (2016))

Parameters	L_{cell}	L_{veh}	v_{max}	T	p_a	p_b	p_c
Units	m	L_{cell}	L_{cell}/s	s	—	—	—
Value	0.5	15	60	1.8	0.85	0.52	0.1
Parameters	a	b_{max}	b_{defense}	g_{safety}	v_c	α	
Units	L_{cell}/s^2	L_{cell}/s^2	L_{cell}/s^2	L_{cell}	L_{cell}/s	s/L_{cell}	
Value	1	7	2	20	30	10	



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 21 The spatiotemporal diagrams of improved two-state model on the circular road. (a) $\rho = 24\text{veh/km}$; (b) $\rho = 38\text{veh/km}$



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 22 The simulation results of spatiotemporal diagram. (a) $\rho = 33\text{veh/km}$; (b) the time series of velocity of a single vehicle

traffic flow. Figure 21 shows that both the $F \rightarrow S$ and the $S \rightarrow J$ transitions can be well depicted. Moreover, the growth of oscillate magnitudes of vehicles in a car platoon are consistent with that reported in the experiment (not shown here, see Tian et al. (2016)). Note that in the improved two-state model, the synchronized flow could coexist with free flow (see Fig. 21a), and traffic oscillations exist in the synchronized flow (see Fig. 22).

Brake Light Models

The class of brake light models also belongs to the generalized NaSch models. Since the status of brake light of preceding car is considered, we review these models separately in this subsection.

The Comfortable Driving (CD) Model

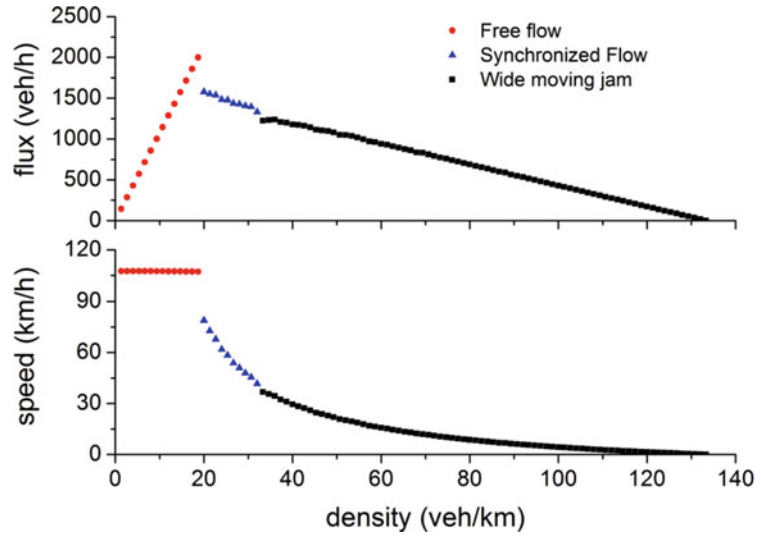
Knospe et al. (2000) proposed a model that considers the desire of the drivers for smooth and comfortable driving. In this model, the formulations of $\hat{a}$, $\hat{d}$, $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ are as follows:

- (i)
$$\hat{a} = \begin{cases} a, & \text{if } (s_l = 0 \text{ and } s = 0) \text{ or } t_h \geq t_s, \\ 0, & \text{otherwise.} \end{cases}$$
- (ii)
$$\hat{d} = d + \max(v_{\text{anti}} - g_{\text{safety}}, 0).$$
- (iii)
$$\hat{v} = v_{\text{max}}$$
- (iv)
$$\hat{v}_r = b$$
- (v)
$$\hat{p} = \begin{cases} p_b, & \text{if } s_l = 1 \text{ and } t_h < t_s, \\ p_0, & \text{if } v_n = 0, \\ p_d, & \text{otherwise.} \end{cases}$$

Here $v_{\text{anti}} = \min(d_l, v_l)$. s and s_l are the status of the brake light of the vehicle and its leader ($s = 1(0)$ means the brake lights is on (off)). $t_h = d/v$ is the time gap. t_s is the safe time headway, which is defined as $t_s = \min(v, h)$. The slow-to-

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 23

The fundamental diagram of CD model on a circular road with length 7500 m. The parameters are shown in Table 9



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 9 Parameter values of CD model. (Taken from Knospe et al. (2000))

Parameters	L_{cell}	L_{veh}	v_{max}	p_b	p_0	p_d	a	b	g_{safety}	h
Units	m	L_{cell}	L_{cell}/s	—	—	—	L_{cell}/s^2	L_{cell}/s^2	L_{cell}	s
Value	1.5	5	20	0.94	0.5	0.1	1	1	7	6

start rule is applied to simulate the wide moving jams in which standing vehicles in the previous time step have a higher randomization probability p_0 than the moving vehicles with p_d . Moreover, $b \geq a$ should be satisfied to simulate the wide moving jams. The brake light state s' in the next time step is determined as follows:

$$s' = \begin{cases} 1, & \text{if } \tilde{v} < v \text{ or } (p = p_b \text{ and } v' < v), \\ 0, & \text{otherwise.} \end{cases}$$

Figure 23 shows the fundamental diagram of the model. Three density ranges can be classified. In the synchronized flow branch, a phase separation phenomenon will occur, and the system is the coexistence of free flow phase and congested flow phase (Fig. 24a). Note that the congested flow consists of both synchronized flow and jams. In other words, the model cannot correctly reproduce the synchronized flow. With the increase of the density, the free flow region shrinks (see Fig. 24a, b). In the wide moving jam branch, the free flow region disappears, and only the congested flow exists (Fig. 24c). If one continues to increase the

density, the jams will gradually invade the synchronized flow (Fig. 24c, d). Unfortunately, the metastable states and the $F \rightarrow S$ and $S \rightarrow J$ transitions cannot be simulated by CD model.

The Modified Comfortable Driving Model

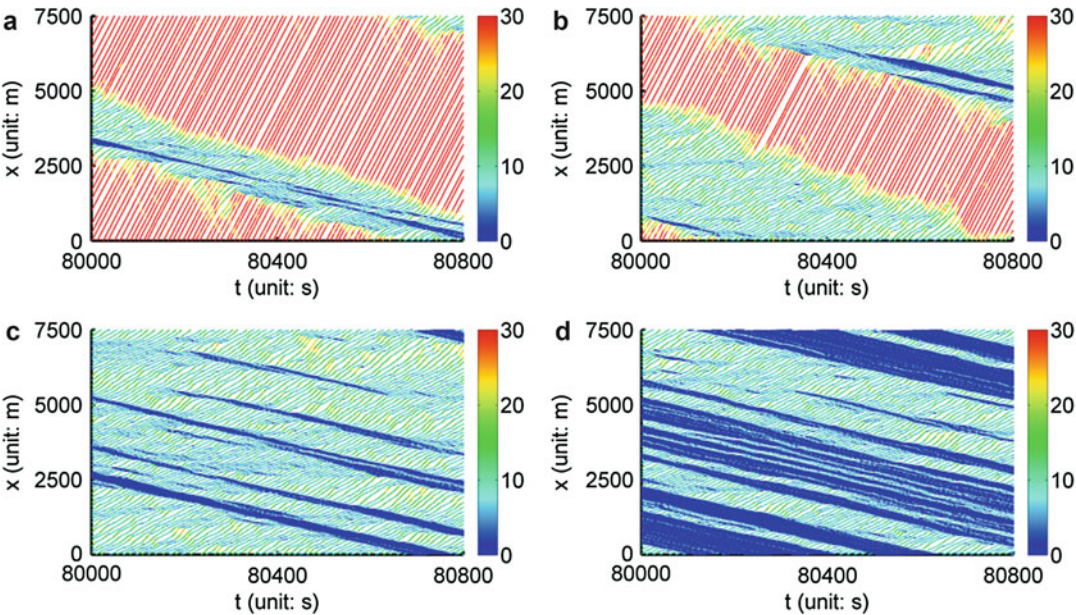
Jiang and Wu (2003) proposed a modified comfortable driving (MCD) model to reproduce both light and heavy synchronized flows. In this model, the formulations of $\hat{a}$, $\hat{d}$, $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ are as follows:

$$(i) \quad \hat{a} = \begin{cases} a_1, & \text{if } (s_l = 0 \text{ or } t_h \geq t_s) \text{ and } v > 0, \\ a_2, & \text{if } v = 0, \\ 0, & \text{otherwise.} \end{cases}$$

$$(ii) \quad \hat{d} = d_n + \max(v_{\text{anti}} - g_{\text{safety}}, 0)$$

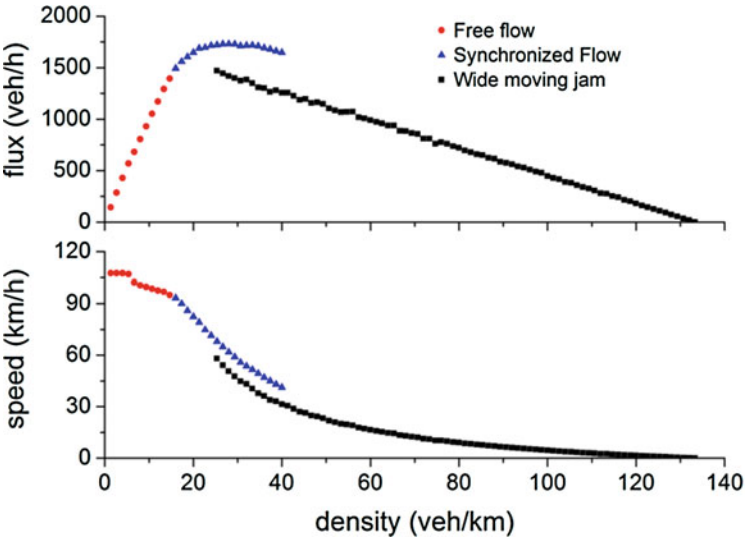
$$(iii) \quad \hat{v} = v_{\text{max}}.$$

$$(iv) \quad \hat{v}_r = b.$$



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 24 The spatiotemporal diagram of the CD model on the circular road. (a) $\rho = 23\text{veh/km}$; (b) $\rho = 28\text{veh/km}$; (c) $\rho = 49\text{veh/km}$; (d) $\rho = 76\text{veh/km}$

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 25 The fundamental diagram of MCD model on the circular road. The parameters are shown in Table 10



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 10 Parameter values of MCD model. (Taken from Jiang and Wu (2003))

Parameters	L_{cell}	L_{veh}	v_{max}	p_b	p_0	p_d	b	a_1	a_2	g_{safety}	h	t_c
Units	m	L_{cell}	L_{cell}/s	—	—	—	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	L_{cell}	s	s
Value	1.5	5	20	0.94	0.5	0.1	1	2	1	7	6	7

$$(v) \quad \hat{p} = \begin{cases} p_b, & \text{if } s_l = 1 \text{ and } t_h < t_s, \\ p_0, & \text{if } v = 0 \text{ and } t_{st} \geq t_c, \\ p_d, & \text{otherwise.} \end{cases} \quad t'_{\text{free}} = \begin{cases} t_{\text{free}} + 1, & \text{if } v' \geq v_c, \\ 0, & \text{otherwise.} \end{cases}$$

In step (i), the acceleration of a stopped car is assumed to be a_2 and that of a moving car is a_1 . Note that sometimes the vehicles do not accelerate in step (i). Moreover, $b \geq a_2$ should be satisfied to simulate the wide moving jams. The brake light state s' in the next time step is determined as follows in the MCD model, which is different from that in the CD model:

$$s' = \begin{cases} 1, & \text{if } v' < v, \\ 0, & \text{if } v' > v, \\ s, & \text{otherwise.} \end{cases}$$

Figure 25 shows the fundamental diagram of the MCD model. One can see that the results are similar to the Gao model. The $F \rightarrow S$ transition cannot be reproduced.

To overcome this deficiency, Jiang and Wu (2005) further introduced a new variable t_{free} . It denotes the time that car n is in the state $v \geq v_c$. It is supposed that if $v' \geq v_c$ and $t_{\text{free}} \geq t_{c1}$, then s remains to be zero despite of the value of v . For the determination of t_{free} , it is simply that

Here v_c and t_{c1} are parameters and t'_{free} is the t_{free} at the next time step.

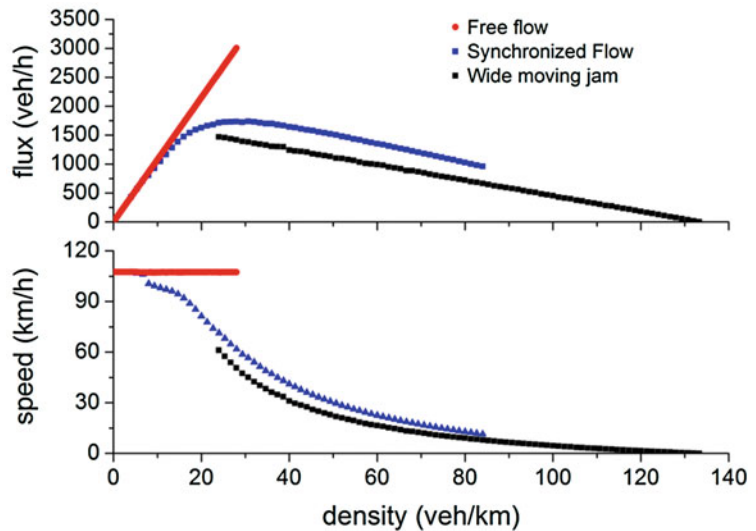
Simulation shows that the improved MCD model can simulate the $F \rightarrow S$ transition (see the fundamental diagram in Fig. 26 and spatiotemporal evolution pattern in Fig. 27a). Note that in the improved MCD model, the synchronized flow does not coexist with the free flow (see Figs. 27a and 28a). Nevertheless, the model has the same deficiency as the improved model of Gao et al., i.e., the velocity fluctuates too much in synchronized flow (see Fig. 28b).

The Desire Time Gap Brake Light (DTGBL) Model Tian et al. (2017) found that in the comfortable driving model, when the leading vehicle decelerates sharply, there is not enough time for following vehicle to smoothly adapt its speed due to the reaction delay. As a result, jam emerges, and synchronized flow cannot be maintained.

To overcome this deficiency, they have revised the braking rule by introducing a desired time gap larger than 1 s. In the new model, the formulations of $\hat{a}$, $\hat{d}$, $\hat{v}$, $\hat{v}_r$, and $\hat{p}$ are as follows:

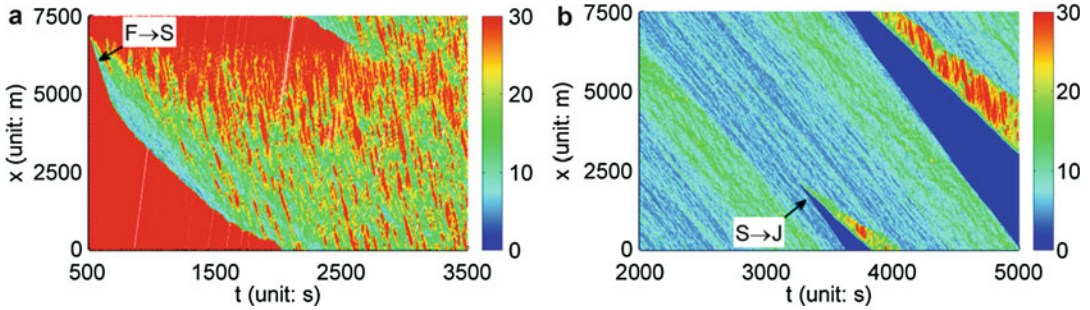
Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 26

The fundamental diagram of improved MCD model on the circular road with length 7500 m. The parameters are shown in Table 11

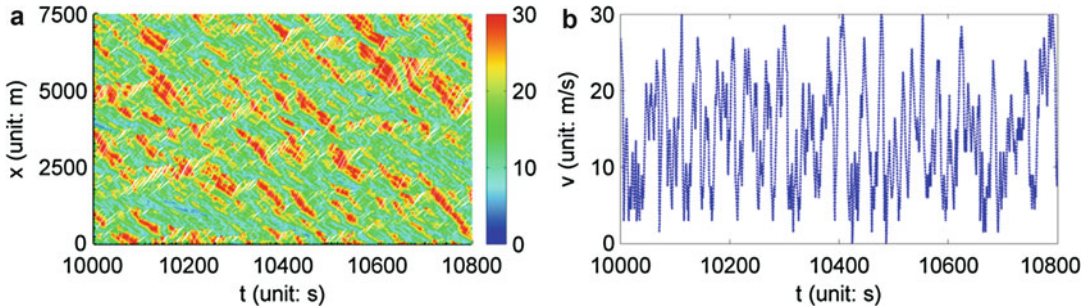


Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 11 Parameter values of improved MCD model. (Taken from Jiang and Wu (2005))

Parameters	L_{cell}	L_{veh}	v_{max}	p_b	p_0	p_d	b
Units	m	L_{cell}	L_{cell}/s	—	—	—	L_{cell}/s^2
Value	1.5	5	20	0.94	0.5	0.1	1
Parameters	a_1	a_2	g_{safety}	h	v_c	t_c	t_{c1}
Units	L_{cell}/s^2	L_{cell}/s^2	L_{cell}	s	L_{cell}/s	s	s
Value	2	1	7	6	18	10	30



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 27 The simulation results of the spatiotemporal diagram. (a) $\rho = 29 \text{ veh/km}$; (b) $\rho = 85 \text{ veh/km}$



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 28 The simulation results of spatiotemporal diagram. (a) $\rho = 33 \text{ veh/km}$; (b) the time series of velocity of a single vehicle

$$(i) \quad \hat{a} = \begin{cases} a_1, & \text{if } (s_l = 0 \text{ or } t_h \geq t_s) \text{ and } v > 0 \\ a_2, & \text{otherwise} \end{cases}$$

$$(ii) \quad \hat{d} = \left\lceil \left(d_n + \max(v_{\text{anti}} - g_{\text{safety}}, 0) \right) / T \right\rceil$$

$$(iii) \quad \hat{v} = v_{\text{max}}$$

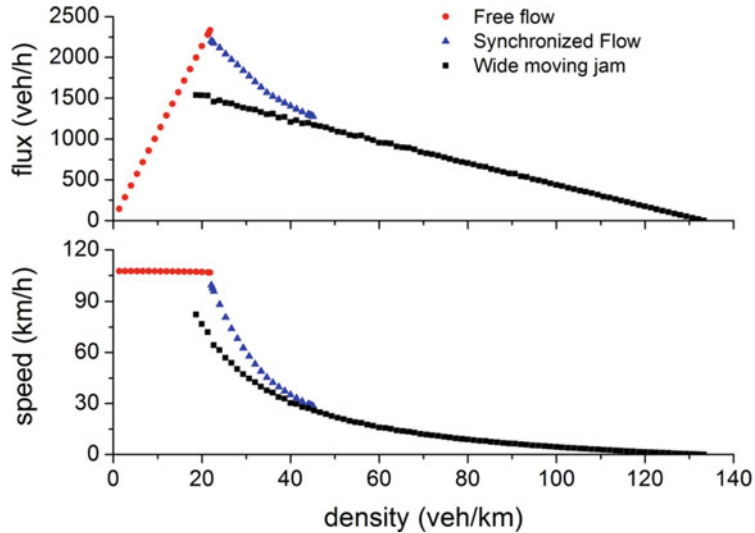
$$(iv) \quad \hat{v}_r = b$$

$$(v) \quad \hat{p} = \begin{cases} p_b, & \text{if } s_l = 1 \text{ and } t_h < t_s, \\ p_0, & \text{if } v = 0, \\ p_d, & \text{otherwise.} \end{cases}$$

where the brake light rule in the CD model is applied. Here $\lceil x \rceil$ returns the minimum integer no smaller than x .

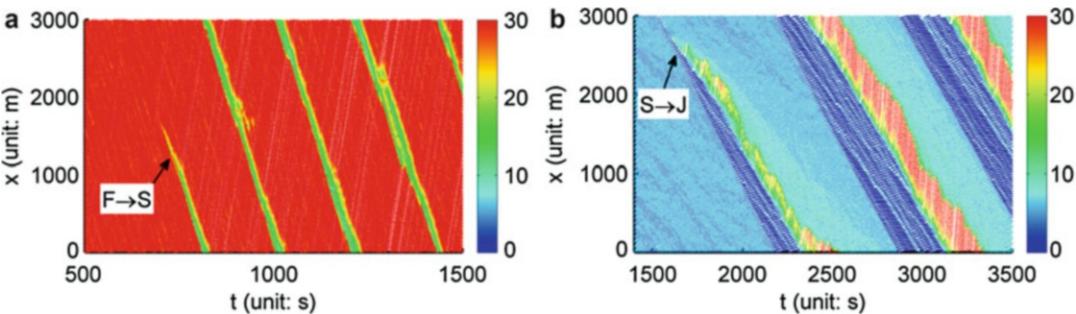
Compared with the CD model, the highlight feature is that the desired time gap is set as $T > 1$. Figure 29 shows the fundamental diagram of

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 29 The fundamental diagram of DTGBL model with length 3000 m. The parameters are shown in Table 12



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 12 Parameter values of DTGBL model. (Taken from Tian (2017))

Parameters	L_{cell}	L_{veh}	v_{max}	T	p_b	p_0	p_d	g_{safety}	h	a_1	a_2	b
Units	m	L_{cell}	L_{cell}/s	s	—	—	—	L_{cell}	s	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$
Value	1.5	5	20	1.8	0.94	0.5	0.1	7	6	2	1	1

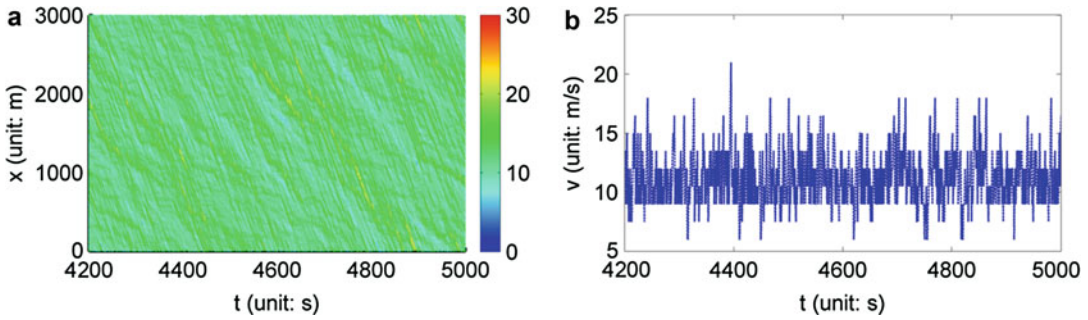


Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 30 The spatiotemporal pattern of (a) $\rho = 22\text{veh/km}$, describes $F \rightarrow S$, (b) $\rho = 67\text{veh/km}$, describes $S \rightarrow J$ transition

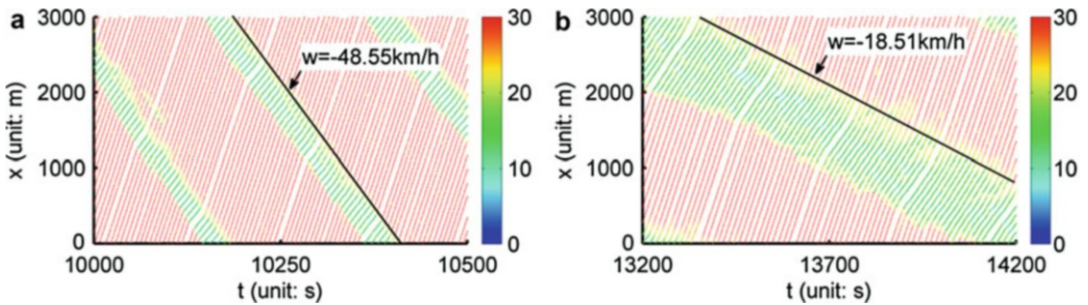
this model, which can reproduce the free flow, the synchronized flow, the jam, as well as the $F \rightarrow S$ (Fig. 30a) and $S \rightarrow J$ transitions (Fig. 30b). Also note that the fluctuation of speed in synchronized flow is reasonable (see Fig. 31).

However, the propagation speed of synchronized flow is too fast in this model (see Fig. 32a).

To overcome this shortcoming, Wang et al. (2017) assumes that (i) the drivers would be sensitive to the brake light only when their speeds are larger than a critical speed v_{cri} and (ii) the effectiveness of the anticipation g_{safety} increases with the increase of the speed of the following vehicle, i.e., $g_{\text{safety}} = \lceil b + g_s v / v_{\text{max}} \rceil$. They have modified the DTGBL model by the reformulations of $\tilde{a}$ and $\tilde{p}$.



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 31 The simulation results of spatiotemporal diagram. (a) $\rho = 37\text{veh/km}$; (b) the time series of velocity of a single vehicle



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 32 The spatiotemporal pattern of (a) DTGBL model with $\rho = 24\text{veh/km}$, (b) the improved DTGBL model with $\rho = 24\text{veh/km}$

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 13 Parameter values of improved DTGBL model. (Taken from Wang et al. (2017))

Parameters	L_{cell}	L_{veh}	v_{max}	T	p_b	p_0	p_d	h	a	b	g_s	v_{cri}
Units	m	L_{cell}	L_{cell}/s	s	—	—	—	s	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	L_{cell}/s
Value	1.5	5	20	2.0	0.94	0.45	0.1	6	1	1	5	9

(i)

$$\hat{a} = a$$

(ii)

$$\hat{p} = \begin{cases} p_b, & \text{if } s_l = 1 \text{ and } t_h < t_{sa} \text{ and } v > v_{cri}, \\ p_0, & \text{if } v = 0, \\ p_d, & \text{otherwise.} \end{cases}$$

The propagation speed of synchronized flow is consistent with the reality (see Fig. 32b). Note that in the DTGBL model, synchronized flow can coexist with the free flow (Table 13).

The Limited Deceleration Model

In the previous models, vehicles have infinite deceleration capability and can decelerate from any speed to zero speed in one time step. In some models, the limited deceleration capability of vehicles has been considered (Lee et al. 2004; Lárraga and Alvarez-Icaza 2010; Chmura et al. 2014; Jin and Wang 2011; Kokubo et al. 2011). One example of this type of models was proposed by Lee et al. (2004). Optimistic and pessimistic driving are introduced. Drivers evaluate the local situation to decide whether they will drive optimistically

($\gamma = 0$) or pessimistically ($\gamma = 1$). The parallel updating rules are as follows:

$$x' = x + v'.$$

Step 1: Determine the safe velocity v_s :

where,

$$v_s = \max \left(c \left| x + \Delta + \sum_{i=0}^{\tau(c)} (c - bi) \leq x' + \sum_{i=1}^{\tau_l(v_l)} (v_l - bi) \right. \right).$$

$$\gamma = \begin{cases} 0, & \text{if } v \leq v_l \leq v_{l+1} \text{ or } v_{l+1} \geq v_{\text{fast}}, \\ 1, & \text{otherwise.} \end{cases}$$

$$\Delta = L_{\text{veh}} + \gamma \max(0, \min(g_{\text{add}}, v - g_{\text{add}})).$$

Step 2: Determine the randomization probability p :

$$\tau(v) = \gamma v/b + (1 - \gamma) \max(0, \min(v/b, t_{\text{safe}}) - 1).$$

$$p = \max(p_d, p_0 - (p_0 - p_d)v/v_{\text{slow}}).$$

Step 3: Velocity update:

$$\tilde{v} = \min(v_{\text{max}}, v + a, \max(0, v - b, v_s)).$$

$$\tau_l(v) = \gamma v/b + (1 - \gamma) \min(v/b, t_{\text{safe}}).$$

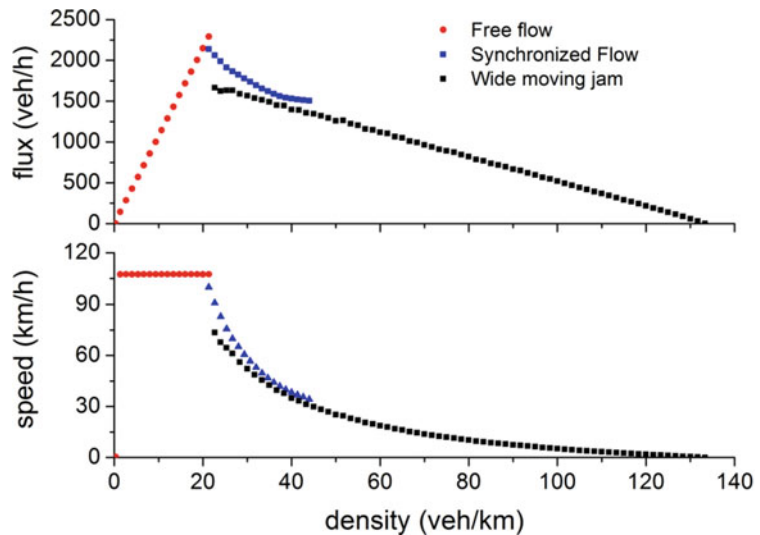
Step 4: Randomization:

$$v' = \begin{cases} \max(0, v - b, \tilde{v} - a), & \text{if } r < p, \\ \max(0, v - b\tilde{v}) & \text{otherwise.} \end{cases}$$

Step 5: Vehicle movement:

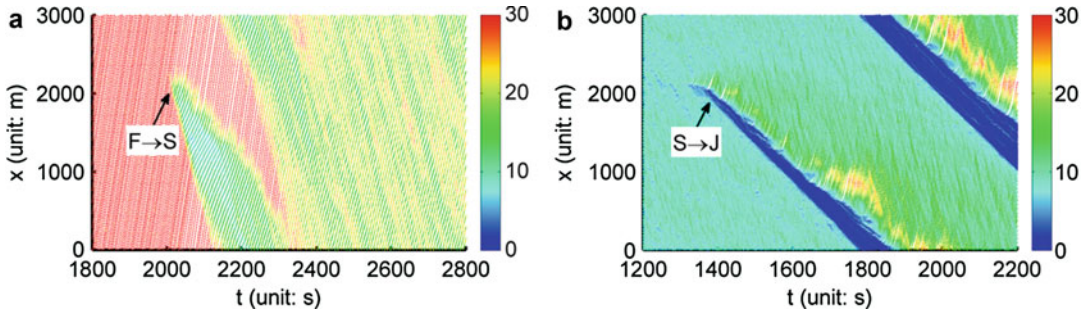
Here, a and b are restricted acceleration and deceleration. γ denotes the driving condition where 0 represents a relative safe condition and 1 represents a relative tense condition. $\tau(v)$ and $\tau_l(v)$, respectively, are the time from deceleration to stop of the followed vehicle and the leading one. c is the safe velocity value of the follower,

Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 33 The fundamental diagram of the Lee model on the circular road with length 3000 m. The parameters are shown in Table 14

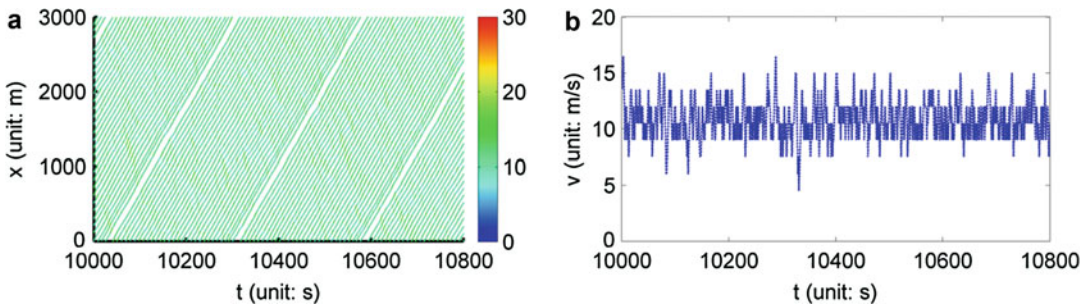


Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Table 14 Parameter values of Lee model. (Taken from Lee et al. (2004))

Parameters	L_{cell}	L_{veh}	v_{max}	a	b	v_{fast}	t_{safe}	g_{add}	p_0	p_d	v_{slow}
Units	m	L_{cell}	L_{cell}/s	$L_{\text{cell}}/\text{s}^2$	$L_{\text{cell}}/\text{s}^2$	L_{cell}/s	s	L_{cell}	—	—	L_{cell}/s
Value	1.5	5	20	1	2	19	3	4	0.32	0.11	5



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 34 The spatiotemporal diagrams of Lee model. (a) $\rho = 27 \text{ veh/km}$; (b) $\rho = 50 \text{ veh/km}$



Cellular Automaton Models in the Framework of Three-Phase Traffic Theory, Fig. 35 The simulation results of spatiotemporal diagram. (a) $\rho = 40 \text{ veh/km}$; (b) the time series of velocity of a single vehicle

and its maximum v_s denotes moving as fast as possible while keeping safe. v_{fast} is a high speed slightly below v_{max} . $v \leq v_l \leq v_{l+1}$ means the drivers drive in a platoon that are driving equally fast or faster than themselves, where v_{l+1} is the speed of next vehicles of their leaders. Δ is the minimal coordinate difference required by the follower to guarantee its safety. g_{add} is introduced for an additional security gap in the defensive state, and t_{safe} is a maximal time step during which the follower observes his/her own safety in the optimistic state.

Figure 33 shows the fundamental diagram, which can reproduce the free flow, the synchronized flow, the jam, as well as the $F \rightarrow S$ (Fig. 34a) and $S \rightarrow J$ transitions (Fig. 34b). Also note that the synchronized flow does not coexist with free flow and the synchronized flow is quite homogeneous (see Fig. 35). Finally, we would like to mention

that due to the limited deceleration capability, this model is not collision-free. Sometimes collision happens.

Conclusion and Outlook

Conclusions

Based on the analysis of the empirical data, Kerner proposed a three-phase theory in the 1990s. In the theory, traffic flow is classified into free flow, synchronized flow, and jam. The transitions in traffic flow usually correspond to $F \rightarrow S$ or $S \rightarrow J$. The direct $F \rightarrow J$ transition seldom occurs. Instead, the two step $F \rightarrow S \rightarrow J$ transition is observed. Different from the linear instability mechanism in two-phase theory, the $F \rightarrow S$ and $S \rightarrow J$ transitions are supposed to be due to over-acceleration effect.

Many models were proposed in the framework of three-phase theory. In the CA models, the randomness is an important component. Thus, the stochasticity is naturally considered, which is found to play a nontrivial role in traffic flow. If the stochasticity is removed from CA models, none of them can show the simulation results discussed in previous sections. Moreover, the time step is usually set to 1 s, while it is usually 0.1 s or even smaller in car-following models. As a result, the simulation time is much less in CA models and suitable for large-scale traffic network.

In this paper, we have reviewed the three-phase CA models developed in the last two decades, which have been roughly classified into a few different types. The models proposed by Kerner and his colleagues took the speed adaptation into account explicitly. Another type of models was developed, considering the impact of brake light of preceding vehicle. The third type is a generalized NaSch model. In the fourth type of models, the limit deceleration capacity of vehicles is introduced.

Future Directions

While all these models can reproduce the synchronized flow, some common deficiencies exist for almost all models. (1) In real traffic flow, with the increase of the density, the free flow branch bends downward. The lowest speed in the free flow can be as low as 70–80 km/h. However, in the CA models, the free flow branch is almost a straight line. The lowest speed in free flow is very close to the maximum speed. Therefore, how to capture the feature of free flow branch is a common task for CA modeling. (2) Three-phase traffic flow models usually have many parameters, and most of them have no direct physical meanings and thus cannot be measured. (3) The rules of many models are usually based on the behavioral assumptions. The data validation and justification are absent. (4) A quantitative comparison with real traffic data is lacked for most CA models. (5) The extensions of the three-phase models to the traffic networks are needed.

Moreover, due to a lack of precise traffic data, there are many other unclear issues, for example, (i) whether the synchronized flow could coexist with free flow, (ii) how does the traffic flow oscillate and what is the oscillation

strength in the synchronized flow, and (iii) how does the free flow spontaneously transit into synchronized flow on a road without bottleneck. To reveal the issues, significant efforts are still needed in the future work toward traffic flow experiment and observation, data collection, and analysis.

Bibliography

- Ahn S, Cassidy MJ (2007) Freeway traffic oscillations and vehicle lane-change maneuvers. In: *Proceedings of the transportation and traffic theory*, 2007
- Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51(2):1035
- Barlovic R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B-Condens Matter Complex Syst* 5(3):793–800
- Benjamin SC, Johnson NF, Hui PM (1996) Cellular automata models of traffic flow along a highway containing a junction. *J Phys A Math Gen* 29(12):3119
- Brackstone M, McDonald M (1999) Car-following: a historical review. *Transport Res F: Traffic Psychol Behav* 2(4):181–196
- Brilon W, Geistefeldt J, Regler M (2005) Reliability of freeway traffic flow: a stochastic concept of capacity. In: *Proceedings of the 16th international symposium on transportation and traffic theory*
- Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6(2):165–184
- Chen D, Laval J, Zheng Z, Ahn S (2012) A behavioral car-following model that captures traffic oscillations. *Transp Res B Methodol* 46(6):744–761
- Chen D, Ahn S, Laval J, Zheng Z (2014) On the periodicity of traffic oscillations and capacity drop: the role of driver characteristics. *Transp Res B Methodol* 59:117–136
- Chmura T, Herz B, Knorr F, Pitz T, Schreckenberg M (2014) A simple stochastic cellular automaton for synchronized traffic flow. *Physica A: Stat Mech Appl* 405:332–337
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329(4–6):199–329
- Chung K, Rudjanakanoknad J, Cassidy MJ (2007) Relation between traffic density and capacity drop at three freeway bottlenecks. *Transp Res B Methodol* 41(1):82–95
- Coifman B, Kim S (2011) Extended bottlenecks, the fundamental relationship, and capacity drop on freeways. *Transp Res A Policy Pract* 45(9):980–991
- Daganzo C, Daganzo CF (1997) *Fundamentals of transportation and traffic operations*, vol 30. Pergamon, Oxford

- Fukui M, Ishibashi Y (1997) Effect of delay in restarting of stopped cars in a one-dimensional traffic model. *J Phys Soc Jpn* 66(2):385–387
- Gao K, Jiang R, Hu SX, Wang BH, Wu QS (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. *Phys Rev E* 76(2):026105
- Gao K, Jiang R, Wang BH, Wu QS (2009) Discontinuous transition from free flow to synchronized flow induced by short-range interaction between vehicles in a three-phase traffic flow model. *Phys A Stat Mech Appl* 388(15–16):3233–3243
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7(4):499–505
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9(4):545–567
- Gipps PG (1981) A behavioural car-following model for computer simulation. *Transp Res B Methodol* 15(2):105–111
- Greenberg H (1959) An analysis of traffic flow. *Oper Res* 7(1):79–85
- Greenshields BD, Channing W, Miller H (1935) A study of traffic capacity. In: Highway research board proceedings, vol 1935. National Research Council (USA), Highway Research Board
- Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Transp Res Rec* 1320:91
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73(4):1067
- Jia B, Li XG, Chen T, Jiang R, Gao ZY (2011) Cellular automaton model with time gap dependent randomisation under Kerner's three-phase traffic theory. *Transportmetrica* 7(2):127–140
- Jiang R, Wu QS (2003) Cellular automata models for synchronized traffic flow. *J Phys A Math Gen* 36(2):381
- Jiang R, Wu QS (2005) First order phase transition from free flow to synchronized flow in a cellular automata model. *Eur Phys J B-Condens Matter Complex Syst* 46(4):581–584
- Jiang R, Wu Q, Zhu Z (2001) Full velocity difference model for a car-following theory. *Phys Rev E* 64(1):017101
- Jiang R, Wu QS, Zhu ZJ (2002) A new continuum model for traffic flow and numerical tests. *Transp Res B Methodol* 36(5):405–419
- Jiang R, Hu MB, Zhang HM, Gao ZY, Jia B, Wu QS, Wang B, Yang M (2014) Traffic experiment reveals the nature of car-following. *PLoS One* 9(4):e94351
- Jiang R, Hu MB, Zhang HM, Gao ZY, Jia B, Wu QS (2015) On some experimental features of car-following behavior and how to model them. *Transp Res B Methodol* 80:338–354
- Jiang R, Jin CJ, Zhang HM, Huang YX, Tian JF, Wang W, Hu MB, Wang H, Jia B (2017) Experimental and empirical investigations of traffic flow instability. *Transp Res Part C: Emerg Technol.* <https://doi.org/10.1016/j.trc.2017.08.024>
- Jin CJ, Wang W (2011) The influence of nonmonotonic synchronized flow branch in a cellular automaton traffic flow model. *Phys A Stat Mech Appl* 390(23–24):4184–4191
- Jin CJ, Wang W, Jiang R, Zhang HM, Wang H, Hu MB (2015) Understanding the structure of hyper-congested traffic from empirical and experimental evidences. *Transp Res Part C: Emerg Technol* 60:324–338
- Kerner BS (1998) Experimental features of self-organization in traffic flow. *Phys Rev Lett* 81(17):3797–3800
- Kerner BS (1999a) Congested traffic flow: observations and theory. *Transp Res Rec J Transp Res Board* 1678(1):160–167
- Kerner BS (1999b) Theory of congested traffic flow: self-organization without bottlenecks. In: 14th international symposium on transportation and traffic theory
- Kerner B (1999c) Congested traffic flow: observations and theory. *Transp Res Rec J Transp Res Board* 1678(1):160–167
- Kerner BS (2000) Experimental features of the emergence of moving jams in free traffic flow. *J Phys A Math Gen* 33(26):L221
- Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65(4):046138
- Kerner BS (2004) The physics of traffic: empirical freeway pattern features, engineering applications, and theory. *Phys Today* 58(11):54–56
- Kerner BS (2009) Introduction to modern traffic flow theory and control: the long road to three-phase traffic theory. Springer, Berlin
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Phys A Stat Mech Appl* 392(21):5261–5282
- Kerner BS (2017) Breakdown in traffic networks: fundamentals of transportation science. Springer, Heidelberg
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35(3):L31
- Kerner BS, Rehborn H (1996) Experimental properties of complexity in traffic flow. *Phys Rev E* 53(5):R4275
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. *Phys Rev Lett* 79(20):4030
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35(47):9971
- Kerner BS, Klenov SL, Schreckenberg M (2011) Simple cellular automaton model for traffic breakdown, highway capacity, and synchronized flow. *Phys Rev E* 84(4):046110
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2000) Towards a realistic microscopic description of highway traffic. *J Phys A Math Gen* 33(48):L477
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. *Phys Rev E* 70(1):016115
- Kokubo S, Tanimoto J, Hagishima A (2011) A new cellular automata model including a decelerating damping

- effect to reproduce Kerner's three-phase theory. *Phys A Stat Mech Appl* 390(4):561–568
- Lárraga ME, Alvarez-Icaza L (2010) Cellular automaton model for traffic flow based on safe driving policies and human reactions. *Phys A Stat Mech Appl* 389(23):5425–5438
- Laval J (2006) Stochastic processes of moving bottlenecks: approximate formulas for highway capacity. *Transp Res Rec: J Transp Res Board* 1988:86–91
- Laval JA, Daganzo CF (2006) Lane-changing in traffic streams. *Transp Res B Methodol* 40(3):251–264
- Laval JA, Leclercq L (2010) A mechanism to describe the formation and propagation of stop-and-go waves in congested freeway traffic. *Philos Trans R Soc Lond A: Math, Phys Eng Sci* 368(1928):4519–4541
- Lee HK, Barlovic R, Schreckenberg M, Kim D (2004) Mechanical restriction versus human overreaction triggering congested traffic states. *Phys Rev Lett* 92(23):238702
- Li X, Ouyang Y (2011) Characterization of traffic oscillation propagation under nonlinear car-following laws. *Transp Res B Methodol* 45(9):1346–1361
- Li X, Wu Q, Jiang R (2001) Cellular automaton model considering the velocity effect of a car on the successive car. *Phys Rev E* 64(6):066128
- Li X, Wang X, Ouyang Y (2012) Prediction and field validation of traffic oscillation propagation under nonlinear car-following laws. *Transp Res B Methodol* 46(3):409–423
- Li X, Cui J, An S, Parsafard M (2014) Stop-and-go traffic analysis: theoretical properties, environmental impacts and oscillation mitigation. *Transp Res B Methodol* 70:319–339
- Lighthill MJ, Whitham GB (1955) On kinematic waves II. A theory of traffic flow on long crowded roads. *Proc R Soc Lond A* 229(1178):317–345. The Royal Society
- Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419(1):1–64
- Mauch M, Cassidy MJ (2002) Freeway traffic oscillations: observations and predictions. In: *Proceedings of the 15th international symposium on transportation and traffic theory*
- May AD (1990) *Traffic flow fundamentals*. Prentice Hall, Englewood Cliffs
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I* 2(12):2221–2229
- Nishinari K, Takahashi D (2000) Multi-value cellular automaton models and metastable states in a congested phase. *J Phys A Math Gen* 33(43):7709
- Payne HJ (1979) FREFLO: a macroscopic simulation model of freeway traffic. *Transp Res Rec* 722:68–77
- Pipes LA (1967) Car following models and the fundamental diagram of road traffic. *Trans Res* 1(1):21–29
- Richards PI (1956) Shock waves on highway. *Oper Res* 4:42–51
- Saberi M, Mahmassani HS (2013) Empirical characterization and interpretation of hysteresis and capacity drop phenomena in freeway networks. *Transp Res Rec: J Transp Res Board*. Transportation Research Board of the National Academies, Washington, DC
- Saifuzzaman M, Zheng Z (2014) Incorporating human factors in car-following models: a review of recent developments and research needs. *Transp Res Part C: Emerg Technol* 48:379–403
- Saifuzzaman M, Zheng Z, Haque MM, Washington S (2017) Understanding the mechanism of traffic hysteresis and traffic oscillations through the change in task difficulty level. *Transp Res B Methodol* 105:523–538
- Schadschneider A, Schreckenberg M (1997) Traffic flow models with 'slow-to-start' rules. *Ann Phys* 509(7):541–551
- Schönhof M, Helbing D (2007) Empirical features of congested traffic states and their implications for traffic modeling. *Transp Sci* 41(2):135–166
- Srivastava A, Geroliminis N (2013) Empirical observations of capacity drop in freeway merges with ramp control and integration in a first-order model. *Transp Res Part C: Emerg Technol* 30:161–177
- Sugiyama Y, Fukui M, Kikuchi M, Hasebe K, Nakayama A, Nishinari K et al (2008) Traffic jams without bottlenecks – experimental evidence for the physical mechanism of the formation of a jam. *New J Phys* 10(3):033001
- Takayasu M, Takayasu H (1993) 1/f noise in a traffic model. *Fractals* 1(04):860–866
- Tian JF, Yuan ZZ, Jia B, Fan HQ, Wang T (2012a) Cellular automaton model in the fundamental diagram approach reproducing the synchronized outflow of wide moving jams. *Phys Lett A* 376(44):2781–2787
- Tian JF, Yuan ZZ, Treiber M, Jia B, Zhang WY (2012b) Cellular automaton model within the fundamental-diagram approach reproducing some findings of the three-phase theory. *Phys A Stat Mech Appl* 391(11):3129–3139
- Tian J, Treiber M, Ma S, Jia B, Zhang W (2015) Microscopic driving theory with oscillatory congested states: model and empirical verification. *Transp Res B Methodol* 71:138–157
- Tian J, Li G, Treiber M, Jiang R, Jia N, Ma S (2016) Cellular automaton model simulating spatiotemporal patterns, phase transitions and concave growth pattern of oscillations in traffic flow. *Trans Res B Methodol* 93:560–575
- Tian J, Jia B, Ma S, Zhu C, Jiang R, Ding Y (2017) Cellular automaton model with dynamical 2D speed-gap relation. *Trans Sci* 51(3):807–822
- Treiber M, Helbing D (2003) Memory effects in microscopic traffic models and wide scattering in flow-density data. *Phys Rev E* 68(4):046119
- Treiber M, Kesting A (2013) *Traffic flow dynamics: data, models and simulation*. no. Book, Whole. Springer, Berlin/Heidelberg
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62(2):1805
- Treiber M, Kesting A, Helbing D (2006) Understanding widely scattered traffic flows, the capacity drop, and platoons as effects of variance-driven time gaps. *Phys Rev E* 74(1):016123

- Treiterer J, Myers J (1974) The hysteresis phenomenon in traffic flow. *Transp Traffic Theory* 6:13–38
- Wang Z, Ma S, Jiang R, Tian J (2017) A cellular automaton model reproducing realistic propagation speed of downstream front of the moving synchronized pattern. *Transportmetrica B: Transp Dyn.* <https://doi.org/10.1080/21680566.2017.1401966>
- Yeo H, Skabardonis A (2009) Understanding stop-and-go traffic in view of asymmetric traffic theory. In: *Transportation and traffic theory 2009: Golden Jubilee*. Springer, Boston, pp 99–115
- Zhao BH, Hu MB, Jiang R, Wu QS (2009) A realistic cellular automaton model for synchronized traffic flow. *Chin Phys Lett* 26(11):118902
- Zheng Z (2014) Recent developments and research needs in modeling lane changing. *Transp Res B Methodol* 60:16–32
- Zheng Z, Ahn S, Chen D, Laval J (2011) Applications of wavelet transform for analysis of freeway traffic: bottlenecks, transient traffic, and traffic oscillations. *Transp Res B Methodol* 45(2):372–384

Autonomous Driving in the Framework of Three-Phase Traffic Theory

Boris S. Kerner
Physics of Transport and Traffic, University
Duisburg-Essen, Duisburg, Germany

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Fundamental Empirical Features of Traffic Breakdown
Classical (Standard) Strategy of Adaptive Cruise Control (ACC)
Failure of Standard Approaches for Simulations of Mixed Traffic Flow
Main Prediction of Three-Phase Traffic Theory
ACC in Framework of Three-Phase Theory (TPACC)
Model of TPACC
Simulations of String Stability of ACC- and TPACC-Vehicles
Speed Disturbances Occurring by Passing of ACC- and TPACC-Vehicles through On-Ramp Bottleneck
Effect of Single Autonomous Driving Vehicle on Probability of Traffic Breakdown
Explanation of Effect of Single Autonomous Driving Vehicle on Speed Disturbance at Bottleneck
Effect of Platoons of Autonomous Driving Vehicles on Probability of Traffic Breakdown in Mixed Traffic Flow
Traffic Stream Characteristics of Mixed Traffic Flow
Discussion
Future Directions
Appendix A. Kerner–Klenov Microscopic Stochastic Traffic Flow Model

Appendix B. Discrete Version of Classical ACC Model

Appendix C. Discrete Version of TPACC Model

Appendix D. Model of On-Ramp Bottleneck

Appendix E. Boundary Conditions for Mixed Traffic Flow

Bibliography

Glossary

Autonomous Driving An autonomous driving vehicle is a self-driving vehicle that can move without a driver. Autonomous driving is realized through the use of an automated system in a vehicle: The automated system has control over the vehicle in traffic flow. For this reason, autonomous driving vehicle is often also called automated driving (or automatic driving) vehicle.

Autonomous Driving in Framework of Three-Phase Traffic Theory An autonomous driving in the framework of the three-phase traffic theory is the autonomous driving for which there is no fixed time headway to the preceding vehicle. This means the existence of an indifference zone in car-following for the autonomous driving vehicle.

Bottleneck Traffic breakdown occurs mostly at road bottlenecks. A road bottleneck can be a result of roadworks, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, traffic signal, etc.

Main Prediction of Three-Phase Traffic Theory The main prediction of the three-phase theory is the *nucleation nature* of an $F \rightarrow S$ transition at a highway bottleneck. The nucleation nature of the $F \rightarrow S$ transition explains the empirical nucleation nature of traffic breakdown at the bottleneck. In its turns, the nucleation nature of the $F \rightarrow S$ transition is governed by an $S \rightarrow F$ instability in synchronized flow. This basic prediction of the three-phase theory has no sense for any other traffic flow theories. For this reason, the three-

phase theory is incommensurable with the other traffic flow theories.

Three-Phase Traffic Theory The three-phase traffic theory (three-phase theory for short) is the framework for understanding of states of empirical traffic flow in three phases: (1) free flow (F), (2) synchronized flow (S), and (3) wide moving jam (J). The synchronized flow and wide moving jam phases belong to congested traffic. An important feature of the three-phase theory is the hypothesis about a two-dimensional (2D) region of states of synchronized flow that means the existence of an indifference zone in car-following in synchronized flow.

Traffic Breakdown In empirical observations, when density in free flow increases and becomes large enough, the phenomenon of the onset of congestion traffic is observed in free flow: The average speed decreases sharply to a lower speed in congested traffic. This speed breakdown observed during the onset of congestion is called the breakdown phenomenon or traffic breakdown. Empirical traffic breakdown at a highway bottleneck is a phase transition from the free flow phase to synchronized flow phase (F→S transition). The empirical traffic breakdown (F→S transition) exhibits the nucleation nature. In the three-phase traffic theory, the empirical nucleation nature of traffic breakdown is explained by the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck.

Definition of the Subject and Its Importance

As generally assumed, future vehicular traffic is a mixed traffic flow consisting of autonomous driving (called also self-driving or automated driving or else automatic driving) vehicles that are randomly distributed between human driving vehicles. Autonomous driving vehicles should increase both traffic safety and capacity of a traffic network. To increase traffic safety, the dynamic behavior of autonomous driving vehicles should be understandable and predictable for drivers of

human driving vehicles in different traffic situations. Otherwise, an autonomous driving vehicle that moves unexpectedly for drivers can be considered as an “obstacle” for the drivers. This can decrease traffic safety. This can also lead to large local disturbances in traffic flow. In its turn, the disturbances can initiate traffic breakdown in traffic network. Thus, autonomous driving systems should be developed in the future that learn from dynamic behavior of human drivers.

In this entry, we make an introduction to an autonomous driving strategy that learns from human behavior. One of the important features of the behavior of drivers is as follows. While following the preceding vehicle, a driver does not try to reach a particular (fixed) time headway to the preceding vehicle. Instead, if the driver cannot pass the preceding vehicle, the driver tries to adapt the vehicle speed to the speed of the preceding vehicle within a time headway range. The assumption of the existence of such an indifference zone for car-following in synchronized flow has firstly made in the three-phase theory (a hypothesis about a two-dimensional (2D) region of synchronized flow states of the three-phase theory). This is one of the reasons why autonomous driving for which there is *no* fixed time headway to the preceding vehicle can be considered autonomous driving in the framework of the three-phase traffic theory. Another reason for this definition is as follows. The classical (standard) autonomous driving vehicle tries to reach a fixed (desired) time headway to the preceding vehicle. The choice of the dynamic strategy for future autonomous driving is of a great importance for traffic safety in mixed traffic flow as well as for the enhancing of network capacity. In this entry, we discuss possible advantages of the autonomous driving strategy in the framework of three-phase traffic theory versus the standard autonomous driving strategy.

Introduction

Future vehicular traffic is a mixed traffic flow consisting of random distributed human driving and autonomous driving vehicles (see, e.g.,

(Bengler et al. 2014; Davis 2004b, 2014, 2016; Ioannou 1997; Ioannou and Chien 1993; Ioannou and Sun 1996; Ioannou and Kosmatopoulos 2000; Levine and Athans 1966; Liang and Peng 1999, 2000; Lin et al. 2009; Martinez and Canudas-de-Wit 2007; Maurer et al. 2015; Meyer and Beiker 2014; Rajamani 2012; Swaroop and Hedrick 1996; Swaroop et al. 2001; Van Brummelen et al. 2018; Varaiya 1993)). There exist a large series of papers by the well-known and massive “Automated Highway System” project involving the US government and a large number of transportation researchers (*Automated Highway Systems*; *Automated Highway Systems*), EU projects (European Roadmap Smart Systems for Automated Driving 2015), and projects made in Germany (*Automatisches Fahren*). A consortium of researches all over the world performed extensive and pioneering research into autonomous and automated driving vehicle systems (see references to these extensive research, e.g., in reviews and books by Ioannou (1997), Ioannou and Sun (1996), Ioannou and Kosmatopoulos (2000), Shladover (1995), Rajamani (2012), Meyer and Beiker (2014), Bengler et al. (2014), and Van Brummelen et al. (2018)).

An autonomous driving vehicle is a self-driving vehicle that can move without a driver. Autonomous driving is realized through the use of an automated system in a vehicle: The automated system has control over the vehicle in traffic flow. For this reason, autonomous driving vehicle is also called automated driving (or automatic driving) vehicle. It should be noted that in the engineering science, the terms *autonomous driving* and *automated driving* are not synonyms. There are two reasons for this. While an autonomous driving vehicle should be able to move without a driver in the vehicle, there are several different levels of automation associated with automated driving. The levels include, for example, a level of “conditional automation” in which the driver must be present to provide any corrections when needed and a level of “full automation” in which the automated vehicle system is in complete control of the vehicle and human presence is no longer needed. Additionally, in contrast with autonomous driving vehicle that moves fully

autonomous from other vehicles, it is often assumed that automated driving can be supported by so-called cooperative driving that can be realized through a diverse variety of cooperative automated systems like vehicle-to-vehicle (V2V) communication (ad hoc vehicle networks) and vehicle-to-infrastructure (V2X) communication. However, to study dynamic strategies for future reliable autonomous driving that should increase network capacity and traffic safety, in this entry we limit a consideration of an automated vehicle system that is in complete control of the vehicle as well as we assume that there are no cooperative vehicle systems that can support automated driving. In other words, for the subject discussed in this entry, there is no difference between the terms *autonomous driving* and *automated driving*.

As mentioned, autonomous driving vehicles should considerably enhance highway capacity. Highway capacity is limited by traffic breakdown at road bottlenecks. Traffic breakdown is a transition from free flow at a bottleneck to congested traffic at the bottleneck (see reviews and books (Bellomo et al. 2002; Brockfeld et al. 2003; Chowdhury et al. 2000; Daganzo 1997; Elefteriadou 2014; Ferrara et al. 2018; Gartner et al. 1997, 2001; Gazis 2002; Haight 1963; Helbing 2001; Highway Capacity Manual 2000, 2010; Kerner 2004, 2009, 2017a; Leutzbach 1988; Mahnke et al. 2005, 2009; May 1990; Nagatani 2002; Nagel et al. 2003; Newell 1982; Papageorgiou 1983; Saifuzzaman and Zheng 2014; Schadschneider et al. 2011; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974). Elefteriadou et al. have found (Elefteriadou et al. 1995) that empirical traffic breakdown at a highway bottleneck exhibits a probabilistic nature: At the same flow rate in free flow at a bottleneck, traffic breakdown can occur, but it should not necessarily occur. Empirical probability of traffic breakdown at highway bottleneck found firstly by Persaud et al. (1998) is a growing function of the flow rate in free flow. Because empirical traffic breakdown in free flow at a bottleneck is the probabilistic phenomenon, the probability of traffic breakdown in free flow at the bottleneck is one of the main characteristics of the traffic stream.

A study of the probability of traffic breakdown in free flow at the bottleneck presented in this review entry allows us to make the following general conclusions:

1. The dynamics of an autonomous driving vehicle should learn from the common features of human driving vehicles in different traffic situations. Indeed, if the dynamics of the autonomous driving vehicle is qualitatively different from the common dynamic behavior of human driving vehicles, an autonomous driving vehicle might be considered an “obstacle” for human driving vehicles in mixed traffic flow. This can reduce traffic safety as well as decrease highway capacity. This explains the necessity of the use of autonomous driving vehicles whose dynamics is familiar for humans. One of the important features of the common dynamic behavior of human driving vehicles found out in empirical data and firstly incorporated in the three-phase traffic theory is an indifference zone in car-following in synchronized flow leading to a two-dimensional (2D) region of synchronized flow states. In this entry, we will consider an autonomous driving in the framework of the three-phase theory introduced in (Kerner 2007a, b; c, 2008) and firstly studied in (Kerner 2017c, 2018b) in which such an indifference zone has been incorporated.
2. To find results of the effect of autonomous driving vehicles on mixed traffic flow that are valid for the reality, the dynamics of human driving vehicles in mixed traffic flow must be simulated with a microscopic traffic flow model in the framework of the three-phase theory. Indeed, we will explain that if one of the standard traffic flow models for simulations of human driving vehicles is used, then results of a study of the effect of autonomous driving vehicles are invalid or even incorrect for the real world. We use the term “standard” for traffic flow models that are related to the state of the art in traffic and transportation research. Examples of standard (classical) traffic flow models can be found in reviews and books (Bellomo et al. 2002; Brockfeld et al. 2003;

Chowdhury et al. 2000; Daganzo 1997; Elefteriadou 2014; Ferrara et al. 2018; Gartner et al. 1997, 2001; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005, 2009; Nagatani 2002; Nagel et al. 2003; Newell 1982; Papageorgiou 1983; Saifuzzaman and Zheng 2014; Schadschneider et al. 2011; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974). This criticism on the standard (classical) traffic flow models has been explained in the review entry in this Encyclopedia (Kerner 2017d).

As mentioned, the main objective of this review entry is to find the effect of different features of the dynamics of autonomous driving vehicles in mixed traffic flow on the probability of traffic breakdown in free flow at a highway bottleneck. We will focus on a mixed traffic flow with a small percentage (denoted by γ) of autonomous driving vehicles. There are the following reasons for this limitation of the study made in this entry. First, in the near future only a very small percentage γ of autonomous driving vehicles in mixed traffic flow can be expected. Second, we would like to study whether already a single autonomous driving vehicle surrounded by human driving vehicles can effect on traffic breakdown at the bottleneck. Indeed, as already explained in the review paper of the Encyclopedia (Kerner 2017b), already a single autonomous driving vehicle whose dynamics is based on the classical (standard) approach can provoke traffic breakdown at the bottleneck. In this entry we will find that such a deterioration of traffic system does not occur when an autonomous driving vehicle learns from driver behavior in car-following as introduced in the three-phase theory.

As known (see, e.g., reviews and books (Elefteriadou 2014; Highway Capacity Manual 2000, 2010; Kerner 2004, 2009, 2017a; May 1990) and references there), most important features of traffic breakdown in free flow at an on-ramp bottleneck on a single-lane road are qualitatively the same as those in highly heterogeneous traffic flow consisting of very different types of vehicles on multilane road with different types of road bottlenecks. In particular, this conclusion is

related to the empirical flow rate dependence of the breakdown probability (Elefteriadou 2014; Kerner 2017a). Therefore, to find the effect of different features of the dynamics of autonomous driving vehicles in mixed traffic flow on the probability of traffic breakdown at a road bottleneck, it is sufficient to study a simple case of mixed vehicular traffic where traffic consists only of two types of vehicles (human driving and autonomous driving vehicles) moving on a single-lane road with an on-ramp bottleneck.

On the single-lane road, no vehicles can pass. For this reason, the effect of autonomous driving on traffic breakdown and highway capacity can be understood through an analysis of an adaptive cruise control (ACC) in a vehicle: An ACC-vehicle follows the preceding vehicle (that can be either a human driving vehicle or an ACC-vehicle) automatically based on some ACC dynamics rules of motion (see, e.g., (Davis 2004b, 2014; Ioannou 1997; Ioannou and Chien 1993; Ioannou and Sun 1996; Ioannou and Kosmatopoulos 2000; Levine and Athans 1966; Liang and Peng 1999, 2000; Rajamani 2012; Swaroop and Hedrick 1996; Swaroop et al. 2001)). In 2004 the author introduced a strategy of ACC in the framework of the three-phase theory (Kerner 2007a, b, c, 2008) called TPACC. One of the most important features of TPACC is the existence of the indifference zone in car-following of the three-phase theory. In this review entry, we will show that the TPACC strategy can exhibit important advantages in comparison with the classical (standard) ACC strategy. Main contributions of this review entry are as follows:

- (i) We explain a simple TPACC model (Kerner 2017c, 2018b) that allows us to simulate physical features of mixed traffic flow consisting of human driving vehicles and TPACC-vehicles.
- (ii) We show that the mean amplitude of speed disturbances at a road bottleneck occurring through TPACC-vehicle can be considerably smaller than introduced by classical (standard) ACC-vehicles at the same model parameters.
- (iii) We show that in mixed traffic flow with TPACC-vehicles, the probability of traffic breakdown at a road bottleneck can be considerably smaller than in mixed traffic flow with classical ACC-vehicles.
- (iv) We explain the physics of the improving of the traffic stream through TPACC-vehicles mentioned in item (iii).

The entry is organized as follows: First, we discuss fundamental empirical features of traffic breakdown and stochastic highway capacity (section “[Fundamental Empirical Features of Traffic Breakdown](#)”). Classical (standard) ACC strategy is the subject of section “[Classical \(Standard\) Strategy of Adaptive Cruise Control \(ACC\)](#).” The reason for the failure of standard approaches for simulations of mixed traffic flow is briefly discussed in section “[Failure of Standard Approaches for Simulations of Mixed Traffic Flow](#).” The main prediction of the three-phase theory is formulated in section “[Main Prediction of Three-Phase Traffic Theory](#).” The strategy to autonomous driving in the framework of the three-phase theory called TPACC is discussed in section “[ACC in Framework of Three-Phase Theory \(TPACC\)](#).” In section “[Model of TPACC](#),” we consider a simple TPACC model that allows us to study the physics of autonomous driving in the framework of the three-phase theory. Simulations of string stability of ACC-vehicles and TPACC-vehicles at an on-ramp bottleneck are made in section “[Simulations of String Stability of ACC- and TPACC-Vehicles](#).” Speed disturbances that occur by passing of ACC-vehicles and TPACC-vehicles through the on-ramp bottleneck are the subject of section “[Speed Disturbances Occurring by Passing of ACC- and TPACC-Vehicles Through On-Ramp Bottleneck](#).” An analysis of the effect of a single autonomous driving vehicle on traffic breakdown in free flow at the bottleneck is presented in section “[Effect of Single Autonomous Driving Vehicle on Probability of Traffic Breakdown](#).” Explanations of the effect of single autonomous driving vehicles on the probability of traffic breakdown in mixed traffic flow are given in section “[Explanation of Effect of Single Autonomous Driving Vehicle on Speed Disturbance at](#)

Bottleneck.” The effect of platoons of autonomous driving vehicles on the probability of traffic breakdown in mixed traffic flow is discussed in section “[Effect of Platoons of Autonomous Driving Vehicles on Probability of Traffic Breakdown in Mixed Traffic Flow](#).” Traffic stream flow characteristics of mixed traffic flow are discussed in section “[Traffic Stream Characteristics of Mixed Traffic Flow](#).” In section “[Discussion](#),” we consider the applicability of the TPACC model for a reliable analysis of some features of future autonomous driving in mixed traffic flow (section “[About Applicability of Model Results for Future Autonomous Driving in Mixed Traffic Flow](#)”) and formulate paper conclusions (section “[Conclusions](#)”). In Appendix A, we present the Kerner–Klenov stochastic microscopic three-phase model for human driving vehicles (Kerner and Klenov 2002, 2003, 2009) used for simulations of mixed traffic flow, in Appendix B we explain simulations of the classical ACC model, in Appendix C we consider a discrete TPACC model used for simulations of TPACC model, in Appendix D we consider a model of vehicle merging at an on-ramp bottleneck, and in Appendix E we consider boundary conditions used for simulations of mixed traffic flow.

Fundamental Empirical Features of Traffic Breakdown

To study the effect of autonomous driving on traffic breakdown and highway capacity, we should discuss the fundamental empirical features of traffic breakdown and highway capacity.

First, we should note that in all empirical data measured over years in different countries, traffic breakdown at a highway bottleneck is a phase transition from free flow (F) to the synchronized flow phase (S) of congested traffic (F → S transition); the F → S transition (traffic breakdown) exhibits the nucleation nature (Kerner 1999a, b, c). This means that traffic breakdown occurs in a metastable free flow with respect to an F→S transition at the bottleneck. Detailed explanations of the synchronized flow phase of congested traffic can be found in the books (Kerner 2004, 2009, 2017a) as well as in this Encyclopedia (Kerner 2017b, d).

This metastability of free flow with respect to the F → S transition is as follows (Kerner 1999a, b, c; Kerner et al. 2015): There can be many speed (density, flow rate) disturbances in free flow at the bottleneck. Amplitudes of the disturbances can be very different. When a disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown occurs. Such a disturbance resulting in the breakdown is called *nucleus* for the breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays. As a result, no traffic breakdown occurs.

As explained in the book (Kerner 2017a), the empirical nucleation nature of traffic breakdown is the empirical fundamental of transportation science. In particular, one of the most important consequences of the empirical nucleation nature of traffic breakdown at a highway bottleneck is as follows (Kerner 2004, 2017a): The metastability of free flow with respect to the F→S transition is realized within a flow rate range

$$C_{\min} \leq q_{\text{sum}} < C_{\max}, \quad (1)$$

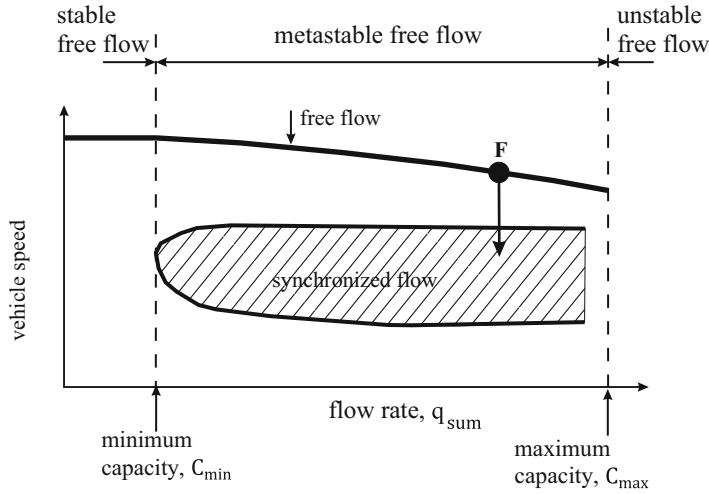
where q_{sum} is the flow rate in free flow at a highway bottleneck, $C_{\min}$ is a minimum highway capacity, and $C_{\max}$ is a maximum highway capacity. Under conditions (1), traffic breakdown can be induced at the bottleneck.

This conclusion of empirical observations of traffic breakdown (F → S transition) at the bottleneck leads to the following definition of stochastic highway capacity of free flow at a highway bottleneck made in the three-phase theory (Kerner 2004):

- *At any time instant*, there are an infinite number of highway capacities C of free flow at the bottleneck. The range of the infinite number of highway capacities is limited by the minimum highway capacity $C_{\min}$ and the maximum highway capacity $C_{\max}$ (Fig. 1):

$$C_{\min} \leq C \leq C_{\max}, \quad (2)$$

where $C_{\min} < C_{\max}$. The existence of an infinite number of highway capacities at any time instant means that highway capacity is stochastic.



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 1 Qualitative explanation of the empirical metastability of free flow with respect to traffic breakdown ($F \rightarrow S$ transition) at bottleneck in three-phase theory (Kerner 1999a, c, 2004, 2017a). Qualitative Z-speed–flow rate characteristic for traffic breakdown:

F, free flow; S, synchronized flow (two-dimensional (2D) hatched region). Arrow $F \rightarrow S$ illustrates symbolically one of possible $F \rightarrow S$ transitions occurring in a metastable state of free flow when a nucleus appears in the free flow at the bottleneck. Adapted from (Kerner 2017a)

The physical sense of this capacity definition is as follows. Highway capacity is limited by traffic breakdown in an initial free flow at a highway bottleneck. In other words, any flow rate q_{sum} in free flow at the bottleneck at which traffic breakdown can occur is highway capacity. At any time instant, there are the infinite number of such highway capacities $C = q_{\text{sum}}$ at which traffic breakdown *can* occur. These capacities satisfy conditions (2). Thus, in the three-phase theory, any flow rate q_{sum} in free flow at a highway bottleneck that satisfies conditions

$$C_{\min} \leq q_{\text{sum}} \leq C_{\max} \quad (3)$$

is equal to one of the stochastic highway capacities of free flow at the bottleneck (Fig. 1).

Figure 1 together with conditions (1) explains the term *metastable free flow with respect to an $F \rightarrow S$ transition (traffic breakdown) at the bottleneck* as follows (labeled by “metastable free flow” in Fig. 1). For each flow rate q_{sum} in free flow at the bottleneck that satisfies conditions (1), there can be a transition from a state of the free flow traffic phase (F) to a state of the synchronized flow

traffic phase (S) at the bottleneck. This $F \rightarrow S$ transition (traffic breakdown) at the bottleneck occurs only, when a nucleus required for the breakdown appears at the bottleneck.

A detailed consideration of the nucleation nature of traffic breakdown and stochastic (probabilistic) features of the infinite number of the highway capacities can be found in this Encyclopedia (Kerner 2017b, d) as well as in the book (Kerner 2017a).

Classical (Standard) Strategy of Adaptive Cruise Control (ACC)

In a classical (standard) ACC model, acceleration (deceleration) $a^{(\text{ACC})}$ of the ACC-vehicle is determined by the space gap to the preceding vehicle g and the relative speed $\Delta v = v_\ell - v$ measured by the ACC-vehicle as well as by a desired time headway $\tau_d^{(\text{ACC})}$ of the ACC-vehicle to the preceding vehicle (see, e.g., (Davis 2004b, 2014; Ioannou and Chien 1993; Levine and Athans 1966; Liang and Peng 1999, 2000; Rajamani 2012; Swaroop and Hedrick 1996; Swaroop et al. 2001) and references there):

$$a^{(\text{ACC})} = K_1 \left(g - v\tau_d^{(\text{ACC})} \right) + K_2 (v_\ell - v), \quad (4)$$

where v is the speed of the ACC-vehicle, and v_ℓ is the speed of the preceding vehicle; here and below v , v_ℓ and g are time functions; K_1 and K_2 are coefficients of ACC adaptation. It is well-known that there can be string instability of a long enough platoon of ACC-vehicles (4) (Davis 2004b, 2014; Ioannou and Chien 1993; Levine and Athans 1966; Liang and Peng 1999, 2000; Rajamani 2012; Swaroop and Hedrick 1996; Swaroop et al. 2001). As found by Liang and Peng (1999), the string instability occurs under condition $K_2 < \left(2 - K_1 \left(\tau_d^{(\text{ACC})} \right)^2 \right) / 2\tau_d^{(\text{ACC})}$. The sense of this condition is as follows. We consider a long enough platoon consisting of ACC-vehicles that dynamics satisfies Eq. (4). We assume that initially ACC-vehicles in the platoon move at a time-independent speed at the desired time headway $\tau_d^{(\text{ACC})}$ to each other. If a local speed (or time headway) disturbance appears in the platoon, the disturbance begins to grow in its amplitude. The growth of the disturbance destroys the steadily ACC-vehicle motion in the platoon.

Coefficients K_2 and K_1 of classical ACC (4) can be chosen to satisfy condition for string stability

$$K_2 > \left(2 - K_1 \left(\tau_d^{(\text{ACC})} \right)^2 \right) / 2\tau_d^{(\text{ACC})} \quad (5)$$

Under condition (5), any small local speed (or time headway) disturbance that appears in the platoon decays over time. Therefore, ACC-vehicles in the platoon remain to move at a time-independent speed at the desired time headway $\tau_d^{(\text{ACC})}$ to each other.

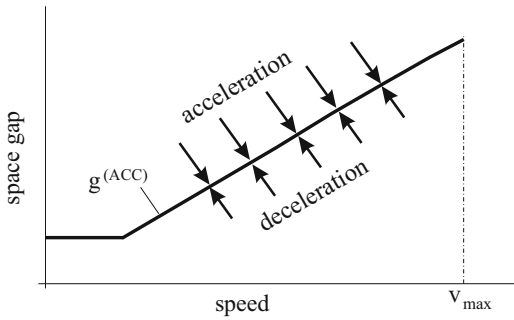
It should be noted that in the works devoted to analysis of the effect of autonomous driving on traffic flow (Bose and Ioannou 2003; Dharba and Rajagopal 1999; Delis et al. 2015; Kesting et al. 2007, 2008, 2010; Kukuchi et al. 2003; Li and Shrivastava 2002; Mamouei et al. 2018; Marsden et al. 2001; Ngoduy 2012, 2013; Ntousakis et al. 2015; Perraki et al. 2018; Roncoli et al. 2015; Shladover et al. 2012; Sharon et al. 2017; Suzuki 2003; Talebpour and Mahmassani 2016; Treiber

and Helbing 2001; Treiber and Kesting 2013; van Arem et al. 2006; VanderWerf et al. 2001, 2002; Wang et al. 2017; Zhou et al. 2017; Zhou and Peng 2005), traffic flow models for human driving vehicles have been used in which at a given time-independent vehicle speed, there is a single model solution for a space gap to the preceding vehicle. Therefore, for such hypothetical steady-state model solutions, there is a one-dimensional (1D) relationship between a chosen speed and the related desired (or optimal) space gap to the preceding vehicle. This well-known assumption for traffic flow models of human driving vehicles used in (Bose and Ioannou 2003; Dharba and Rajagopal 1999; Delis et al. 2015; Kesting et al. 2007, 2008, 2010; Kukuchi et al. 2003; Li and Shrivastava 2002; Mamouei et al. 2018; Marsden et al. 2001; Ngoduy 2012, 2013; Ntousakis et al. 2015; Perraki et al. 2018; Roncoli et al. 2015; Shladover et al. 2012; Sharon et al. 2017; Suzuki 2003; Talebpour and Mahmassani 2016; Treiber and Helbing 2001; Treiber and Kesting 2013; van Arem et al. 2006; VanderWerf et al. 2001, 2002; Wang et al. 2017; Zhou et al. 2017; Zhou and Peng 2005) is qualitatively the same as the existence of a desired time headway $\tau_d^{(\text{ACC})}$ of the ACC-vehicle to the preceding vehicle: For the classical ACC rule (4) that satisfies condition for string stability (5), at a given ACC speed v , there is a single operating point for a desired space gap (Fig. 2)

$$g^{(\text{ACC})} = v\tau_d^{(\text{ACC})} \quad (6)$$

Failure of Standard Approaches for Simulations of Mixed Traffic Flow

It should be mentioned that the effect of classical ACC-vehicles (4) on mixed traffic flow consisting of ACC-vehicles and human driving vehicle was intensively considered already in the 1990s–2000s in the works by Dharba and Rajagopal (1999), Marsden et al. (2001), VanderWerf et al. (2001, 2002), Treiber and Helbing (2001), Li and Shrivastava (2002), Kukuchi et al. (2003), Bose and Ioannou (2003), Suzuki (2003), Davis (2004b),



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 2 Operating points of the classical model of ACC (4) under condition of string stability (5); Qualitative speed dependence of desired space gap $g^{(ACC)} = v\tau_d^{(ACC)}$ (6) (see, e.g., (Davis 2004b, 2014; Ioannou and Chien 1993; Levine and Athans 1966; Liang and Peng 1999, 2000; Rajamani 2012; Swaroop and Hedrick 1996; Swaroop et al. 2001))

Zhou and Peng (2005), van Arem et al. (2006), and Kesting et al. (2007, 2008, 2010); this is a subject of intensive studies up to now (see, e.g., (Chen et al. 2017; Davis 2014; Delis et al. 2015; Han and Ahn 2018; Mamouei et al. 2018; Ngoduy 2012, 2013; Ntousakis et al. 2015; Perraki et al. 2018; Roncoli et al. 2015; Shladover et al. 2012; Sharon et al. 2017; Talebpour and Mahmassani 2016; Treiber and Kesting 2013; Wang et al. 2017; Zhou et al. 2017) and references there).

As proven in details in the books (Kerner 2004, 2009, 2017a), classical (standard) traffic flow theories and models of human driving vehicles cannot explain the empirical nucleation nature of traffic breakdown (F→S transition) at the bottleneck.

- Traffic and transportation theories, which are not consistent with the nucleation nature of traffic breakdown (F→S transition) at highway bottlenecks, cannot be applied for the development of reliable traffic control, dynamic traffic assignment, as well as other reliable ITS applications in traffic and transportation networks.

This criticism on all standard traffic flow models has been explained in details in the book (Kerner 2017a); a brief explanation of this critical statement can also be found in the Encyclopedia (Kerner 2017d).

In almost all studies of mixed traffic flow that the author knows, models of human driving vehicles have been used that cannot explain the nucleation nature of traffic breakdown (F→S transition) at the bottleneck. This criticism is also related to traffic flow theories and models used for studies of human driving vehicles in mixed traffic flow in Refs. (Bose and Ioannou 2003; Chen et al. 2017; Dharba and Rajagopal 1999; Delis et al. 2015; Han and Ahn 2018; Kesting et al. 2007, 2008, 2010; Kukuchi et al. 2003; Li and Shrivastava 2002; Mamouei et al. 2018; Marsden et al. 2001; Ngoduy 2012, 2013; Ntousakis et al. 2015; Perraki et al. 2018; Roncoli et al. 2015; Shladover et al. 2012; Sharon et al. 2017; Suzuki 2003; Talebpour and Mahmassani 2016; Treiber and Helbing 2001; Treiber and Kesting 2013; van Arem et al. 2006; VanderWerf et al. 2001, 2002; Wang et al. 2017; Zhou et al. 2017; Zhou and Peng 2005). Because the standard traffic flow theories and models contradict the empirical fundamental of transportation science, we made the following conclusion:

- Applications of standard traffic flow theories and models of human driving vehicles used, for example, in the studies of mixed traffic flow in (Bose and Ioannou 2003; Chen et al. 2017; Dharba and Rajagopal 1999; Delis et al. 2015; Han and Ahn 2018; Kesting et al. 2007, 2008, 2010; Kukuchi et al. 2003; Li and Shrivastava 2002; Mamouei et al. 2018; Marsden et al. 2001; Ngoduy 2012, 2013; Ntousakis et al. 2015; Perraki et al. 2018; Roncoli et al. 2015; Shladover et al. 2012; Sharon et al. 2017; Suzuki 2003; Talebpour and Mahmassani 2016; Treiber and Helbing 2001; Treiber and Kesting 2013; van Arem et al. 2006; VanderWerf et al. 2001, 2002; Wang et al. 2017; Zhou et al. 2017; Zhou and Peng 2005) lead to invalid (and sometime incorrect) conclusions about the effect of autonomous driving vehicles on highway capacity and traffic breakdown in mixed traffic flow.

For this reason, in this entry we will not consider results of (Bose and Ioannou 2003; Chen et al. 2017; Dharba and Rajagopal 1999; Delis et al. 2015; Han and Ahn 2018; Kesting et al.

2007, 2008, 2010; Kukuchi et al. 2003; Li and Shrivastava 2002; Mamouei et al. 2018; Marsden et al. 2001; Ngoduy 2012, 2013; Ntousakis et al. 2015; Perraki et al. 2018; Roncoli et al. 2015; Shladover et al. 2012; Sharon et al. 2017; Suzuki 2003; Talebpour and Mahmassani 2016; Treiber and Helbing 2001; Treiber and Kesting 2013; van Arem et al. 2006; VanderWerf et al. 2001, 2002; Wang et al. 2017; Zhou et al. 2017; Zhou and Peng 2005) as well as results of all other simulations in which standard traffic flow models for the description of human driving vehicles in mixed traffic flow have been used.

Main Prediction of Three-Phase Traffic Theory

To explain the empirical nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a bottleneck, in 1990s the author introduced the three-phase traffic theory (Kerner 1998b, c, 1999a, b, c). The three-phase theory is the framework for understanding of states of empirical traffic flow in three phases:

1. Free flow (F)
2. Synchronized flow (S)
3. Wide moving jam (J)

The synchronized flow and wide moving jam phases belong to congested traffic. A detailed consideration of features of the three phases F, S, and J of the three-phase theory can be found in the books and reviews (Kerner 2004, 2009, 2013, 2016, 2017a) as well as in this Encyclopedia (Kerner 2017b, d).

The main prediction of the three-phase theory is as follows.

- There is an $S \rightarrow F$ instability in synchronized flow. The $S \rightarrow F$ instability is a growing wave of a local increase in the speed in synchronized flow. The uninterrupted growth of this wave leads to a transition from synchronized flow to free flow. The $S \rightarrow F$ instability exhibits the nucleation nature: Only a large enough initial local increase in the speed in synchronized flow can lead to the $S \rightarrow F$ instability, whereas

local disturbances of small enough local speed increase in synchronized flow decay.

- The nucleation nature of the $S \rightarrow F$ instability governs the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck. In its turn, the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck explains the empirical nucleation nature of traffic breakdown that is the empirical fundamental of transportation science (Kerner 2004, 2015a).

The nucleation nature of the $S \rightarrow F$ instability and the associated metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck have no sense for the classical (standard) traffic and transportation theories. Therefore, the three-phase theory is incommensurable with all earlier traffic flow theories and models (for a review, see the book (Kerner 2017a) as well as this Encyclopedia (Kerner 2017b, d)). The term “incommensurable” has been introduced by Kuhn (2012) to explain a paradigm shift in a scientific field.

One of the first traffic flow models in the framework of the three-phase theory is the Kerner–Klenov microscopic stochastic model (Kerner and Klenov 2002, 2003, 2009). This traffic flow model can show the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck as observed in empirical data.

The effect of classical ACC-vehicles (4) on the probability of traffic breakdown in mixed traffic flow in which human driving vehicles have been simulated with the Kerner–Klenov model in the framework of the three-phase theory has been discussed in (Kerner 2016). The results of (Kerner 2016) have already been considered in this Encyclopedia (Kerner 2017b). It has been shown that even if a platoon of classical ACC-vehicles (4) is stable (5), it can occur that already a small share of classical ACC-vehicles in mixed traffic flow can deteriorate traffic while provoking traffic breakdown at network bottlenecks (Kerner 2016).

In this Encyclopedia entry, we consider a model of ACC in the framework of the three-phase theory (called TPACC, see explanations below) introduced in (Kerner 2017c, 2018b). Based on the model of TPACC, we make a comparison of the effect of classical ACC-vehicles and TPACC-vehicles on the probability of traffic breakdown at a road

bottleneck in mixed traffic flow. Because the Kerner–Klenov microscopic stochastic model (Kerner and Klenov 2002, 2003, 2009) can show the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at the bottleneck as observed in empirical data, for all simulations of human driving vehicles in mixed traffic flow, we use the Kerner–Klenov traffic flow model (see Appendix A).

ACC in Framework of Three-Phase Theory (TPACC)

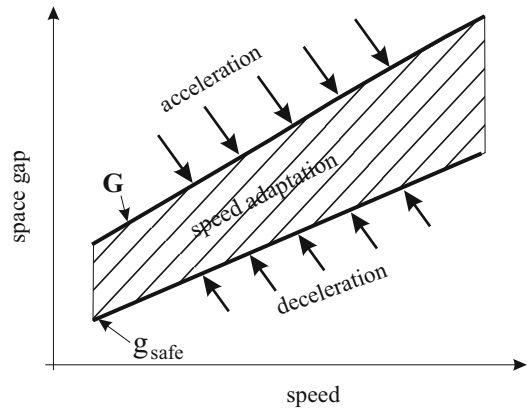
A study of real field traffic data (Kerner 1998b, c, 1999a, b, c, 2000, 2001, 2002a, b; Kerner and Rehborn 1996, 1997) shows that the existence of a desired time headway $\tau_d^{(ACC)}$ of the ACC-vehicle to the preceding vehicle is inconsistent with a basic behavior of real drivers in car-following: Empirical data shows that real drivers do not try to reach a fixed time headway to the preceding vehicle in car-following.

To explain this empirical fact, the author has introduced a hypothesis about the existence of a two-dimensional (2D) region of synchronized flow states (Kerner 1998b, c, 1999a, b, c): In the three-phase theory, it is assumed that when a driver approaches a slower moving preceding vehicle and the driver cannot pass it, the driver decelerates within a synchronization space gap G . This speed adaptation to the speed of the preceding vehicle occurs without caring what the precise space gap g to the preceding vehicle is as long as it is not smaller than a safe space gap g_{safe} (Kerner 1998b, c, 1999a, b, c). The speed adaptation occurring within the synchronization space gap G leads to a 2D region of synchronized flow states (dashed region in Fig. 3) determined by conditions

$$g_{safe} \leq g \leq G. \quad (7)$$

The 2D region of synchronized flow states (dashed region in Fig. 3) can also be considered “indifference zone” in car-following.

According to (7), drivers do not try to reach a particular (desired or optimal) time headway to the preceding vehicle but adapt the speed while keeping time headway $\tau^{(net)} = g/v$ in a range $\tau_{safe} \leq \tau^{(net)} \leq \tau_G$, where $\tau_G = G/v$, τ_G is a



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 3 Qualitative explanation of indifference zone in car-following for TPACC – ACC in the framework of the three-phase theory. Qualitative presentation of a part of 2D region for operating points of TPACC in the space gap–speed plane: At a given speed of TPACC- vehicle, there is the infinity of operating points of TPACC (Kerner 2007a, b, c, 2008). G is a synchronization space gap; g_{safe} is a safe space gap

synchronization time headway, and $\tau_{safe} = g_{safe}/v$ is a safe time headway, and it is assumed that the speed $v > 0$.

We define “autonomous driving in the framework of three-phase traffic theory” as follows:

- An autonomous driving in the framework of the three-phase traffic theory is the autonomous driving for which there is *no* fixed time headway to the preceding vehicle.

A relation of this definition to real autonomous driving vehicles will be discussed in section “About Applicability of Model Results for Future Autonomous Driving in Mixed Traffic Flow.”

In inventions (Kerner 2007a, b, c, 2008), we have assumed that to satisfy these empirical features of real traffic, acceleration (deceleration) of autonomous driving in the framework of the three-phase theory (for short, three-traffic-phase ACC – TPACC) should be given by formula (Kerner 2007a, b, c, 2008)

$$a^{(TPACC)} = K_{\Delta v}(v_\ell - v) \quad \text{at } g_{safe} \leq g \leq G, \quad (8)$$

where $K_{\Delta v}$ is a dynamic coefficient ($K_{\Delta v} > 0$).

The main reason for the use of the word “three-phase” in the ACC strategy is as follows. A 2D region of operating points of TPACC, i.e., indifference zones in car-following shown by dashed region in Fig. 3, follows from the driver behavior firstly incorporated in the three-phase theory in which a 2D region of steady states of synchronized flow is assumed (Kerner 1998a, b, c, 1999a, b, c). The word “three-phase” for TPACC should emphasize both a qualitative difference between the two types of ACCs and the fact that the idea of TPACC with no fixed time headway to the preceding vehicle has been taken from the three-phase theory.

Both the hypothesis about 2D region of traffic flow (indifference zone in car-following; dashed region in Fig. 3) introduced in the three-phase theory (Kerner 1998b, c, 1999a, b, c), car-following models for human driving vehicles with indifference zones in car-following firstly developed in (Kerner and Klenov 2002), and the concept of TPACC strategy (8) (Kerner 2007a, b, c, 2008) have already been reviewed in other entries of this Encyclopedia (Kerner 2017b, d). However, the TPACC model that is suitable for studies of physical characteristics of TPACC in mixed traffic flow has been derived only recently (Kerner 2017c, 2018b). In this review, based on (Kerner 2017c, 2018b) we discuss the physical characteristics of TPACC-vehicles and the physics of the effect of TPACC-vehicles on traffic breakdown in mixed traffic flow.

Model of TPACC

Following (Kerner 2017c, 2018b), we study the following TPACC model:

$$a^{(\text{TPACC})} = \begin{cases} K_{\Delta v}(v_{\ell} - v) & \text{at } g \leq G \\ K_1(g - v\tau_p) + K_2(v_{\ell} - v) & \text{at } g > G, \end{cases} \quad (9)$$

where it is assumed that

$$g \geq g_{\text{safe}}, \quad (10)$$

τ_p is a model parameter that satisfies condition

$$\tau_p < \tau_G. \quad (11)$$

In comparison with the TPACC-strategy (8), the TPACC model (9) allows us to simulate physical features of TPACC-vehicles in mixed traffic flow.

Under condition (10), from the TPACC model (9), it follows that when the space gap $g(t)$ of the TPACC-vehicle to the preceding vehicle is within the range

$$g_{\text{safe}}(t) \leq g(t) \leq G(t), \quad (12)$$

the acceleration (deceleration) of the TPACC-vehicle does not depend on the space gap $g(t)$. Conditions (12) determine the indifference zone in the space gap for TPACC-vehicle in the car-following process. Because time headway of TPACC-vehicle to the preceding vehicle is given by formula $\tau^{(\text{net})} = g/v$, conditions (12) are equivalent to conditions

$$\tau_{\text{safe}}(t) \leq \tau^{(\text{net})}(t) \leq \tau_G \quad (13)$$

that determine the indifference zone in time headway for TPACC-vehicle in the car-following process. In (13), it is assumed that $v > 0$.

As abovementioned, simulations of human driving vehicles in mixed traffic flow are made in this entry with the Kerner–Klenov stochastic traffic flow model in the framework of the three-phase theory (Kerner and Klenov 2002, 2003). In this model, a discrete time $t = n\tau$, where $n = 0, 1, 2, \dots$; $\tau = 1$ s is time step, is used. The model of human driving vehicles of Refs. (Kerner and Klenov 2002, 2003) is continuous in space. We use a version of this model (Kerner and Klenov 2009) that is discrete in space: A very small value of the discretization space interval $\delta x = 0.01$ m is used in the model. As explained in (Kerner and Klenov 2009), this allows us to make more accurate simulations of traffic breakdown at road bottlenecks. Because the model for human driving vehicles (Kerner and Klenov 2002, 2003, 2009) is discrete in time, we simulate TPACC model (9) with discrete time $t = n\tau$. For the simplicity of the consideration of the effect of either classical ACC-vehicles or TPACC-vehicles on traffic

breakdown in mixed traffic flow, discrete models of human driving vehicles, classical ACC-vehicles, and TPACC-vehicles used in simulations have been given in Appendixes A, B, C, D, and E.

Simulations of String Stability of ACC- and TPACC-Vehicles

First, we consider classical ACC that parameters do *not* satisfy condition for string stability (5). Simulations of string instability of ACC-vehicles on a single-lane road with an on-ramp bottleneck are shown in Fig. 4 (a). In simulations, traffic flow consists of $\gamma = 100\%$ of ACC-vehicles. At the on-ramp bottleneck, the on-ramp inflow with the rate q_{on} and upstream flow with the rate q_{in} merge. This ACC-vehicle merging at the bottleneck causes local speed disturbances at the bottleneck. Because in the simulations presented in Fig. 4a traffic flow consists of 100% ACC-vehicles that do not satisfy condition (5) for string stability, local disturbances grow in the amplitude. Due to this string instability of ACC-vehicles, moving traffic jams emerge spontaneously upstream of the bottleneck (Fig. 4a).

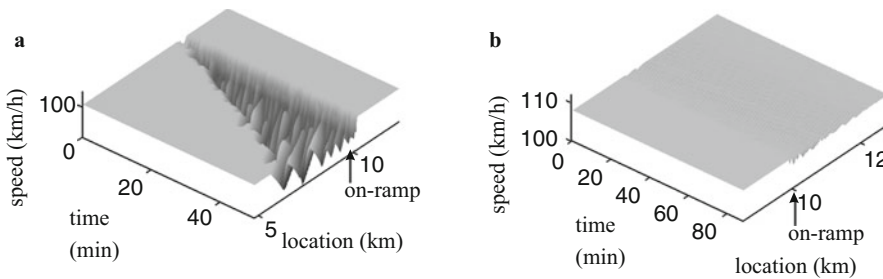
Contrarily to the unstable platoon of classical ACC-vehicles, at the same set of the flow rates q_{on} and q_{in} as well as the same other model

parameters, *no* string instability of any platoon of the TPACC-vehicles is realized: In Fig. 4b, all speed disturbances occurring at the bottleneck decay upstream of the bottleneck. It turns out that as long as time headway between TPACC-vehicles is within the range (13), speed disturbances decay over time. This is because within this range the acceleration (deceleration) of TPACC-vehicles does not depend on time headway to the preceding vehicle.

Speed Disturbances Occurring by Passing of ACC- and TPACC-Vehicles through On-Ramp Bottleneck

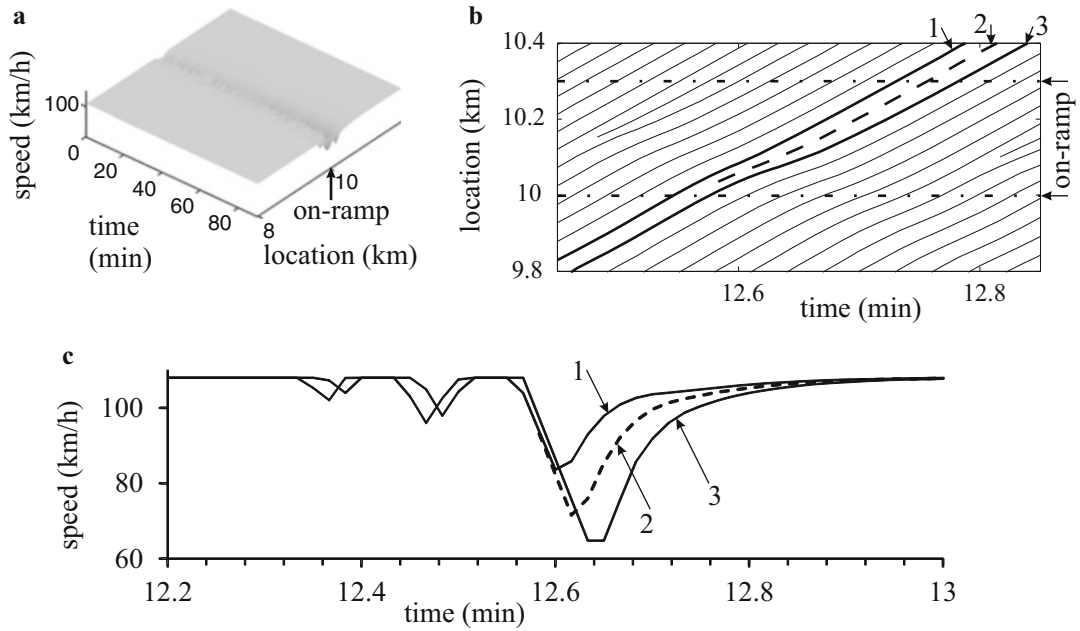
Now we consider ACC-vehicles (4) whose parameters satisfy condition for string stability (5). Indeed, at a larger value of K_2 in (4) as well as at the same desired time headway $\tau_d^{(ACC)} = 1.3$ s and the same set of the flow rates q_{on} and q_{in} as those in Fig. 4 platoons of ACC-vehicles become stable (Fig. 5a).

The objective of our analysis is a comparison of the amplitude of speed disturbances that appear at a highway bottleneck in traffic flow consisting of $\gamma = 100\%$ of autonomous driving vehicles (Fig. 5 for ACC-vehicles and Fig. 6 for TPACC-vehicles).



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 4 Simulations of string instability of the classical ACC model (4) (see discrete ACC model used in simulations in Appendix B) (a) and string stability of TPACC model (9) (see discrete TPACC model used in simulations in Appendix C) (b) in traffic flow consisting of 100% of autonomous driving vehicles on a single-lane road with on-ramp bottleneck located at $x_{on} = 10$ km (see Appendix D): Speed in space and time. Simulation parameters of ACC (a) and TPACC (b) are

identical ones: the on-ramp inflow rate $q_{on} = 320$ vehicles/h and the flow rate upstream of the bottleneck $q_{in} = 2002.6$ vehicles/h, $\tau^{(ACC)} = \tau_p = 1.3$ s, $\tau_G = 1.4$ s, $K_1 = 0.3$ s⁻², and $K_2 = K_{Av} = 0.3$ s⁻¹; $a_{max} = b_{max} = 3$ m/s², $v_{free} = 30$ m/s (108 km/h), vehicle length $d = 7.5$ m. In accordance with the desired time headway $\tau_d^{(ACC)} = 1.3$ s of the ACC-vehicles, the sum flow rate $q_{sum} = q_{in} + q_{on} = 2322.6$ vehicles/h is related to time headway 1.3 s between vehicles in free flow



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 5 Speed disturbances at on-ramp bottleneck in stable free flow with 100% autonomous driving vehicles for classical ACC model (4) (see discrete ACC model used in simulations in Appendix B). (a) Speed in space and time. (b) Fragments of vehicle

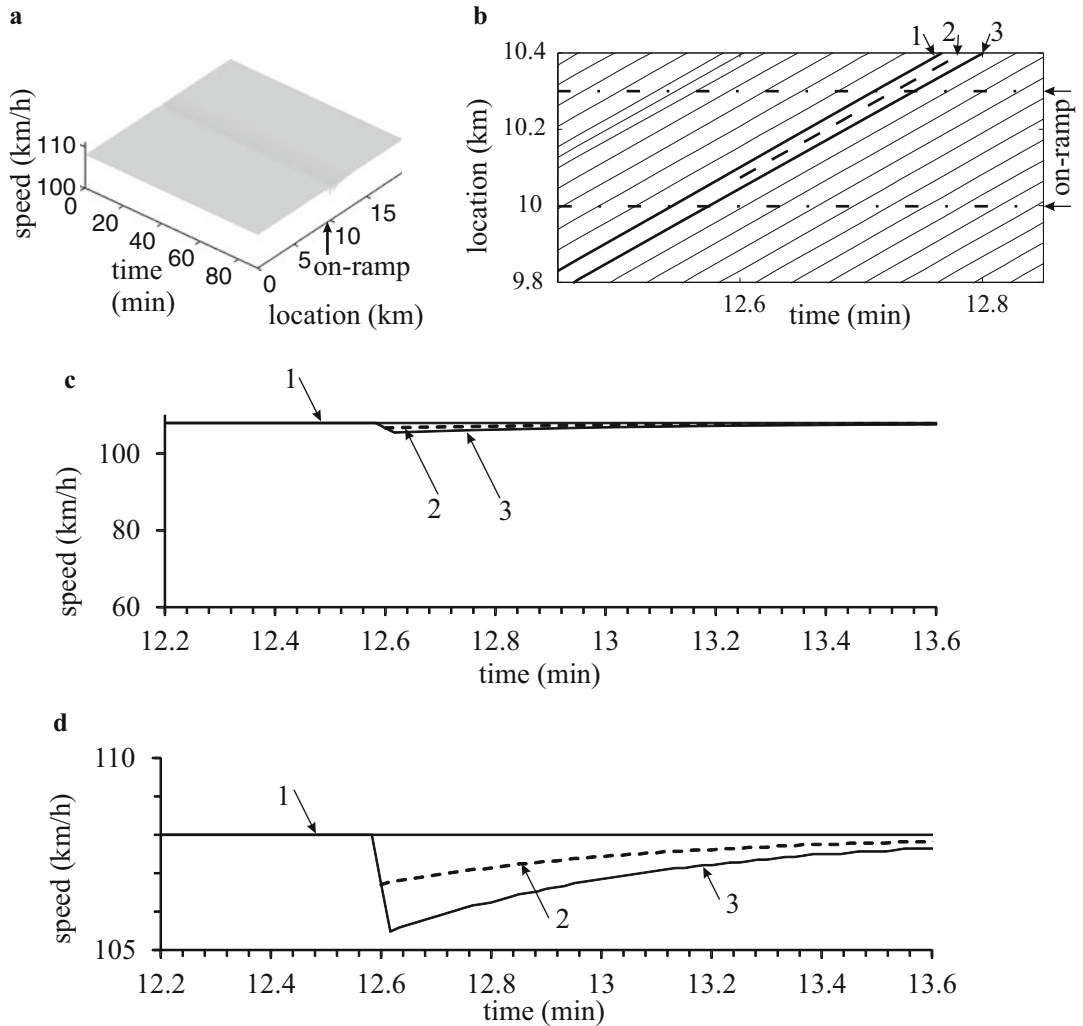
trajectories; (b) is related to (a). (c) Microscopic speeds along vehicle trajectories shown by, respectively, the same numbers in (b). $K_2 = 0.6 \text{ s}^{-1}$. In (b), the on-ramp merging region that is within road locations $x_{\text{on}} \leq x \leq x_{\text{on}}^{(e)}$ (see Appendix D) is labeled by “on-ramp.” Other model parameters are the same as those in Fig. 4

We have found that one of the most important advantages of TPACC strategy (8) in comparison with classical ACC strategy (4) is as follows. At the same model parameters of TPACC-vehicles and ACC-vehicles, TPACC-vehicles (9) cause considerably *smaller* speed disturbances at the bottleneck (Fig. 6) than speed disturbances that appear at the bottleneck through merging of classical ACC-vehicles (4) (Fig. 5).

Indeed, it turns out that local speed disturbances of a considerably large amplitude appear at the bottleneck in traffic flow consisting of $\gamma = 100\%$ of classical ACC-vehicles (4). This case is shown in Fig. 5b, c in which ACC-vehicle 2 merges from the on-ramp onto the main road following ACC-vehicle 1 moving on the main road. To satisfy the desired time headway $\tau_d^{(\text{ACC})}$, ACC-vehicle 2 should decelerate to a lower speed than the minimum speed of ACC-vehicle 1. This deceleration of ACC-vehicle 2 forces the following ACC-vehicle 3 to decelerate while approaching ACC-vehicle 2. Simulations show

that the occurrence of large local speed disturbances at the bottleneck is a basic problem of ACC-vehicles based on the classical ACC strategy (4) in which ACC-vehicles try to reach a desired time headway $\tau_d^{(\text{ACC})}$.

Contrarily to classical ACC-vehicles that cause large local speed disturbances at the bottleneck (Fig. 5b, c), only small local speed disturbances occur in traffic flow consisting of $\gamma = 100\%$ of TPACC-vehicles (Fig. 6). This is because within the range of time headway (13), the acceleration (deceleration) of a TPACC-vehicle does not depend on time headway. This explains small amplitudes of local speed disturbances caused by TPACC-vehicles 2 and 3 at the bottleneck (Fig. 6b–d). The local speed disturbances caused by TPACC-vehicles are so small that in the same speed scale as that used for the classical ACC (Fig. 5c) they cannot almost be resolved in Fig. 6c. Only at considerably larger speed scale the local speed disturbances become visible (Fig. 6d).



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 6 Speed disturbances at on-ramp bottleneck in stable free flow with 100% autonomous driving vehicles for TPACC model (9) (see discrete TPACC model used in simulations in Appendix C). (a) Speed in space and time. (b) Fragment of vehicle trajectories; (b) is related to (a). (c, d) Microscopic speeds along vehicle trajectories shown by, respectively, the same

numbers in (b); (c, d) are related to (b); (c) and (d) show the same speed in different speed scales. Dynamic coefficient K_2 of ACC in Fig. 5 is equal to the related parameter of TPACC: $K_{\Delta v} = 0.6 \text{ s}^{-1}$. Other model parameters are the same as those in Fig. 4. As in Fig. 5b, in (b) the on-ramp merging region that is within road locations $x_{\text{on}} \leq x \leq x_{\text{on}}^{(e)}$ (see Appendix D) is labeled by “on-ramp”

Effect of Single Autonomous Driving Vehicle on Probability of Traffic Breakdown

In the near future, we could expect mixed traffic flow in which the share of autonomous driving vehicles will be very small. Therefore, now we

consider mixed traffic flow consisting of a small percentage $\gamma = 2\%$ of autonomous driving vehicles that are randomly distributed between human driving vehicles. At a such small share of autonomous driving vehicles in mixed traffic flow, a probability that a platoon of several autonomous driving vehicles can occur in mixed traffic flow is

negligible. In other words, almost any autonomous driving vehicle in mixed traffic flow can be considered as a single autonomous driving vehicle surrounded by human driving vehicles. The objective of our study is to show that already a single autonomous driving vehicle can effect considerably on the probability of traffic breakdown in mixed traffic flow at the bottleneck.

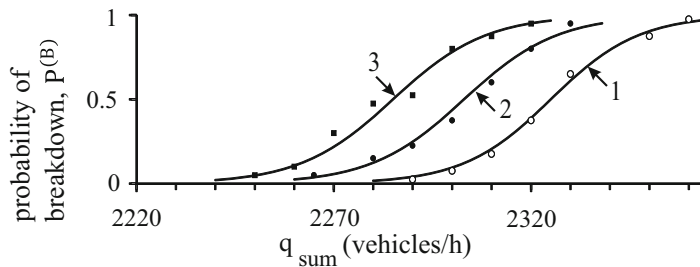
As mentioned in sections “[Fundamental Empirical Features of Traffic Breakdown](#)” and “[Main Prediction of Three-Phase Traffic Theory](#),” traffic breakdown in traffic flow of human driving vehicles is an $F \rightarrow S$ transition. Traffic breakdown occurs in a metastable free flow with respect to the $F \rightarrow S$ transition. Local speed disturbances caused by vehicle interactions in a neighborhood of the bottleneck can randomly initiate traffic breakdown in the metastable free flow. Such traffic breakdown has been called *spontaneous* traffic breakdown (spontaneous $F \rightarrow S$ transition). The larger the amplitude of local speed disturbances at the bottleneck, the more probable the nucleus occurrence for the spontaneous breakdown, i.e., the larger the probability of traffic breakdown $P^{(B)}$ at the bottleneck (Kerner 2004, 2009, 2016, 2017a, b, 2018a).

For calculation of the probability $P^{(B)}(q_{\text{sum}})$ of traffic breakdown in free flow at the bottleneck, at each given value of the flow rate downstream of the bottleneck $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$, different simulation realizations (runs) $N_r = 40$ during the same

time interval for the observing of traffic flow $T_{\text{ob}} = 30$ min have been made (Fig. 7). The different realizations have been performed at the same set of model parameters, however, at different values of the initial values of random function $\text{rand}()$ in the traffic flow model (see Appendix A). Then, $P^{(B)}(q_{\text{sum}}) = n_r/N_r$, where n_r is the number of realizations in which traffic breakdown has occurred during the time interval T_{ob} (more detailed explanations of the calculation of the flow rate function $P^{(B)}(q_{\text{sum}})$ have been given in the book (Kerner 2017a)).

Single TPACC-vehicles moving in such mixed traffic flow cause very small speed disturbances at the bottleneck (Fig. 8a–c). Indeed, we have found that probability of traffic breakdown remains in this mixed flow the same as that in traffic flow consisting of human drivers only (curve 1 in Fig. 7). Thus, single TPACC-vehicles do not effect on the probability of traffic breakdown at the bottleneck.

Contrarily to TPACC-vehicles, single classical ACC-vehicles effect considerably on the probability of traffic breakdown at the bottleneck: The probability of traffic breakdown can increase even when $\gamma = 2\%$ of classical ACC-vehicles is in mixed traffic flow (curves 2 and 3 in Fig. 7). This is because already a single ACC-vehicle can initiate traffic breakdown at the bottleneck (Fig. 9a, b). This deterioration of traffic flow through classical autonomous driving is explained



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 7 Effect of a single autonomous driving vehicle on the probability $P^{(B)}(q_{\text{sum}})$ of traffic breakdown at on-ramp bottleneck on single-lane road in mixed traffic flow. Flow rate functions of breakdown probability $P^{(B)}(q_{\text{sum}})$ (curves 1–3) are calculated through the change in the on-ramp inflow rate q_{on} at a given flow rate $q_{\text{in}} = 2000$ vehicles/h. Curve 1 is related to traffic flow

without autonomous driving vehicles as well as to mixed traffic flow with 2% of TPACC-vehicles. Curves 2 and 3 are related to mixed traffic flow with 2% of ACC-vehicles, respectively, with $\tau_d^{(\text{ACC})} = 1.3$ s (curve 2) and 1.6 s (curve 3). Other model parameters for ACC-vehicles and TPACC-vehicles are, respectively, the same as those in Figs. 5 and 6, respectively

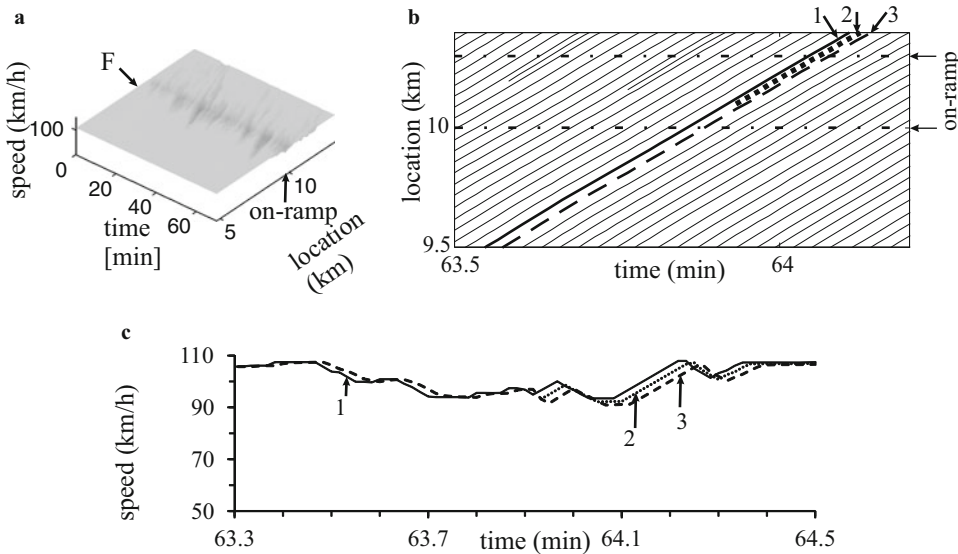
by the occurrence of a large amplitude speed disturbance caused by a classical ACC-vehicle at the bottleneck (dashed vehicle trajectory 3 in Fig. 9b, c).

Explanation of Effect of Single Autonomous Driving Vehicle on Speed Disturbance at Bottleneck

The occurrence of very different amplitudes of local speed disturbances caused by the propagation of a single classical ACC- and TPACC-vehicles through the bottleneck (compare dashed vehicle trajectories 3 in Figs. 8b, c and 9b, c) can be understood if we consider time functions of the space gap $g(t)$ and the acceleration (deceleration) of the ACC- and TPACC-vehicles (Fig. 10). We can see that after the space gap $g(t)$ due to merging of vehicle 2 decreases abruptly (labeled by “merging of vehicle 2” in Fig. 10a, c), both ACC-vehicle and TPACC-vehicle decelerate to increase the space gap $g(t)$. However, the deceleration of

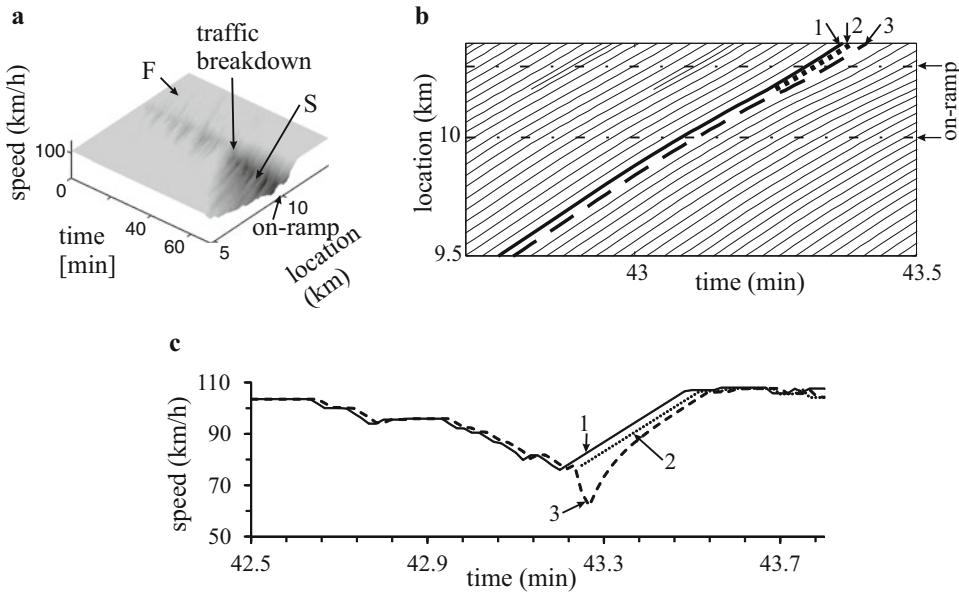
ACC-vehicle (Fig. 10b) is considerably stronger than that of TPACC-vehicle (Fig. 10d). The strong deceleration of the ACC-vehicle explains the strong speed reduction of the ACC-vehicle shown in trajectory 3 in Fig. 9c.

To understand this very different microscopic dynamic behavior of ACC- and TPACC- vehicles, we should mention that in the models of ACC- and TPACC-vehicles (see Appendixes B and C), there is a safe speed v_s that depends on the speed of the preceding vehicle v_A and on the space gap $g(t)$ between vehicles: When the vehicle speed of an autonomous driving vehicle satisfies condition $v(t) \leq v_s(t)$, then the acceleration (deceleration) of autonomous driving vehicles is determined by formula (4) for ACC and formula (9) for TPACC. It should be noted that condition $v(t) \leq v_s(t)$ is equivalent to condition $g(t) \geq g_{\text{safe}}(t)$ for the space gap. For this reason, when the space gap $g(t) < g_{\text{safe}}(t)$, then the deceleration of an autonomous driving vehicle should be at least as strong as it follows from the formulation for the safe speed (see Appendix A). The formulation for



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 8 Explanation of the result shown by curve 1 of Fig. 7 that a single TPACC-vehicle does not effect on the probability of traffic breakdown at on-ramp bottleneck in mixed traffic flow. Speed disturbances occurring at on-ramp bottleneck through a single TPACC-vehicle: (a) speed in space and time, (b) fragment

of vehicle trajectories, (c) microscopic speeds along vehicle trajectories shown by the same numbers in (b), respectively. In (b, c), vehicles 1 and 2 are human driving vehicles, whereas vehicle 3 is TPACC-vehicle. Mixed traffic flow with 2% of TPACC-vehicles; $q_{\text{in}} = 2000$ vehicles/h, $q_{\text{on}} = 280$ vehicles/h. Other model parameters are the same as those in Fig. 6. In (a), F, free flow



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 9 Explanation of the effect of a single classical ACC-vehicle on the probability of traffic breakdown at on-ramp bottleneck in mixed traffic flow associated with curve 2 of Fig. 7. Speed disturbances occurring at on-ramp bottleneck through a single ACC-vehicle (a–c): (a) speed in space and time, (b) fragment of vehicle trajectories, (c) microscopic speeds along vehicle

trajectories shown by the same numbers in (b), respectively. In (b, c), vehicles 1 and 2 are human driving vehicles, whereas vehicle 3 is ACC-vehicle. Mixed traffic flow with 2% of ACC-vehicles; $q_{in} = 2000$ vehicles/h, $q_{on} = 280$ vehicles/h. Other model parameters for ACC-vehicles are the same as those in Fig. 5. In (a), F, free flow; S, synchronized flow

the safe speed is the same for human driving vehicles, ACC-vehicles, and TPACC-vehicles. The formulation for the safe speed $v_s(v_\ell, g)$ guarantees a collision-less motion of autonomous driving vehicles.

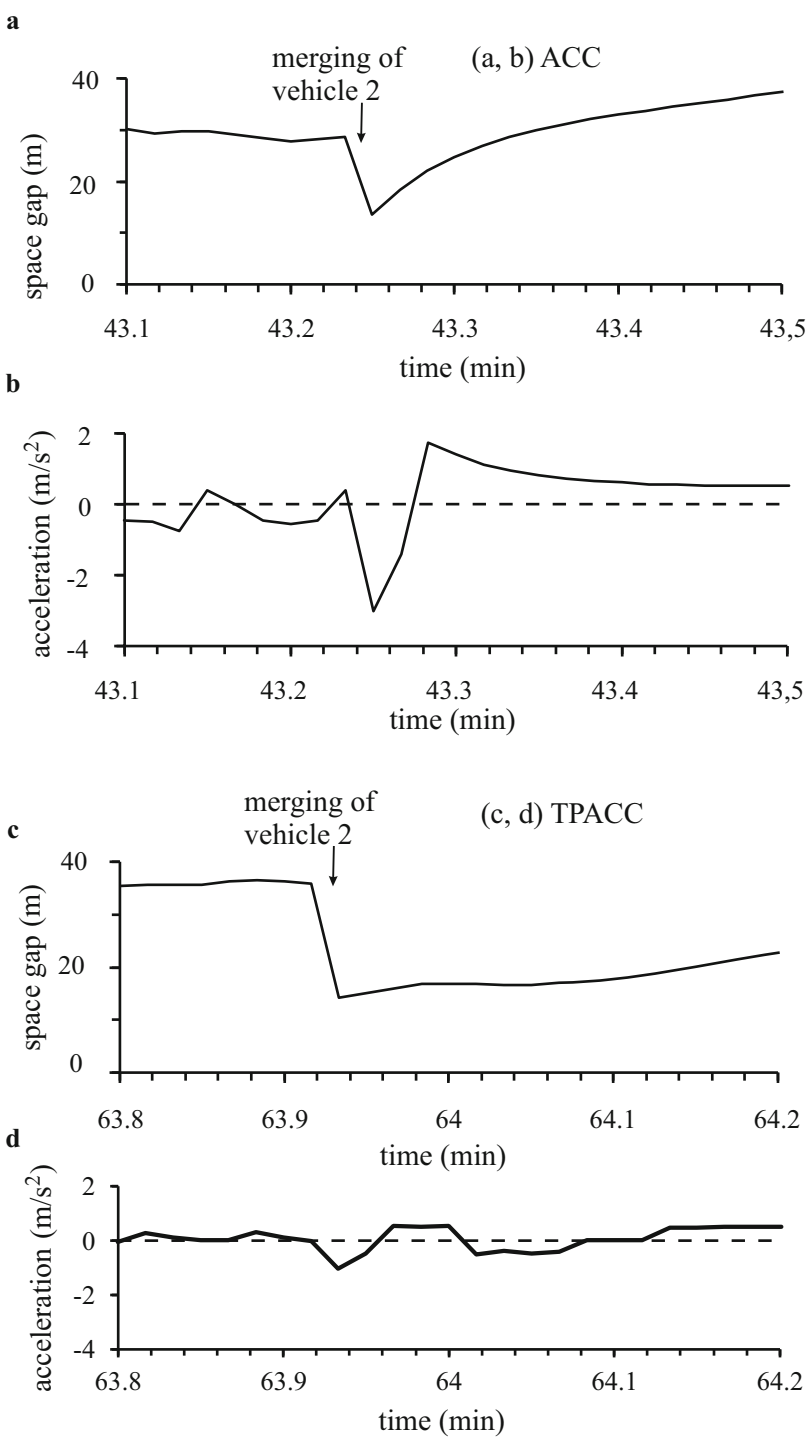
It turns out that although the space gap $g(t)$ due to merging of vehicle 2 decreases abruptly (labeled by “merging of vehicle 2” in Fig. 10a), the deceleration of ACC-vehicle determined by formula (4) leads to a stronger vehicle deceleration (Fig. 11a) than it follows from the formulation for the safe speed $v_s(v_\ell, g)$. Therefore, the stronger vehicle deceleration determined by formula (4) is applied. Due to the merging of vehicle 2 from the on-ramp, there are two effects that follow each other: (i) the abrupt decrease in the space gap $g(t)$ and (ii) the subsequent increase in the speed difference $\Delta v = v_\ell - v$. Because the effect (i) occurs earlier, firstly the deceleration of ACC-vehicle is determined by the term $K_1(g - v\tau_d^{(ACC)})$ in (4) (curve 1 in Fig. 11a).

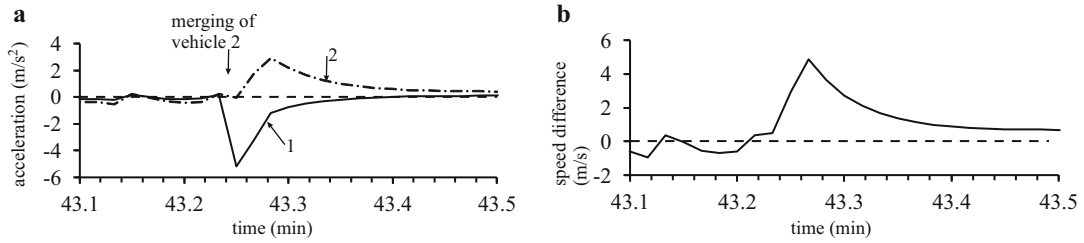
The subsequent increase in the speed difference $\Delta v = v_\ell - v$ that is responsible for the positive value of the term $K_2(v_\ell - v)$ in (4) (curve 2 in Fig. 11a) cannot prevent the strong deceleration of ACC-vehicle mentioned above (Fig. 10b).

Microscopic behavior of TPACC-vehicle (Fig. 12) is qualitatively different from that of ACC-vehicle (Fig. 11) discussed above. Before vehicle 2 merges onto the main road, the space gap $g(t)$ of TPACC-vehicle has satisfied condition (12) of the indifference zone (first gray region in Fig. 12a that is related to time interval before labeling “merging of vehicle 2”). After the space gap $g(t)$ due to merging of vehicle 2 decreases abruptly (labeled by “merging of vehicle 2” in Fig. 12a), condition $g(t) < g_{safe}(t)$ is satisfied for TPACC-vehicle. In contrast with ACC-vehicle, the deceleration of TPACC-vehicle determined by formula (9) leads to a weaker vehicle deceleration than it follows from the formulation for the safe speed $v_s(v_\ell, g)$. As a result, TPACC-vehicle

Autonomous Driving in the Framework of Three-Phase Traffic Theory,

Fig. 10 Explanations of speed disturbances caused by propagation of a single autonomous driving vehicle through bottleneck. **(a, b)** Time functions of the space gap $g(t)$ **(a)** and the acceleration (deceleration) of the ACC-vehicle **(b)** related to vehicle trajectory 3 in Fig. 9b, c. **(c, d)** Time functions of the space gap $g(t)$ **(c)** and the acceleration (deceleration) of the TPACC-vehicle **(d)** related to vehicle trajectory 3 in Fig. 8b, c





Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 11 Continuation of Fig. 10a, b. (a) Components of formula (4) as time functions; solid curve 1 is time dependence of the term $K_1(g - v\tau_d^{(ACC)})$, dashed-dotted curve 2 is time dependence of the term $K_2(v_\ell - v)$. (b) Relative speed between the preceding vehicle and ACC-vehicle. Note

that the deceleration calculated from $K_1(g - v\tau_d^{(ACC)})$ in (4) reaches a large negative value about -5.2 m/s^2 . It should be noted that as long as condition $v(t) \leq v_s(t)$ is satisfied, in the ACC- and TPACC models, there is a limitation of the deceleration by the value -3 m/s^2 . This explains the maximum deceleration of ACC-vehicle -3 m/s^2 shown in Fig. 10b

decelerates in accordance with the formulation of the safe speed $v_s(v_\ell, g)$.

As mentioned, the formulation of the safe speed $v_s(v_\ell, g)$ is related to the behavior of a human driving vehicle (Appendix A). For this reason, the safe speed $v_s(v_\ell, g)$ depends considerably stronger on the speed difference $\Delta v = v_\ell - v$ than on the space gap $g(t)$. When under condition $g(t) < g_{\text{safe}}(t)$ the speed difference $\Delta v = v_\ell - v \geq 0$, the deceleration of the TPACC-vehicle associated with the safety conditions (Figs. 10d and 12c) is considerably smaller than the deceleration of the ACC-vehicle given by formula (4) (Figs. 10b and 11a). Moreover, when the value $\Delta v > 0$ is large enough, condition $g(t) > g_{\text{safe}}(t)$ for the indifference zone is satisfied (second gray region in Fig. 12a). The time evolution of the speed difference $\Delta v(t)$ (Fig. 12b) determines in the large degree the deceleration (acceleration) of TPACC-vehicle (Fig. 12c). Over time, there is an alternation of indifference zones (12) (in which formulas (9) are valid) and safety deceleration regions (in which the TPACC-vehicle moves in accordance with the formulation for the safe speed) (Fig. 12a). This alternation determines a slowly increase in the space gap $g(t)$ of TPACC-vehicle (Fig. 12a). Finally, TPACC-vehicle moves in the indifference zone only.

Both safety conditions of TPACC-vehicle and car-following behavior (9) at the bottleneck (trajectory 3 in Fig. 8c) are similar as those for human driving vehicles (trajectories 1 and 2 in

Fig. 8c). For this reason, we can consider autonomous driving in the framework of the three-phase theory “autonomous driving learning from real driving behavior.”

In particular, when due to the vehicle merging the space gap $g(t)$ becomes shorter than the safe space gap g_{safe} and the speed difference $\Delta v = v_\ell - v \approx 0$, a TPACC-vehicle decelerates as slowly as a human driving vehicle does. This explains why in simulations of mixed traffic flow with TPACC-vehicles presented in Fig. 8b, c no large speed disturbances occur at the bottleneck. This explains also why in contrast with classical autonomous driving (curve 2 in Fig. 7) single TPACC-vehicles do not effect on the probability of traffic breakdown at the bottleneck in mixed traffic flow (curve 1 in Fig. 7).

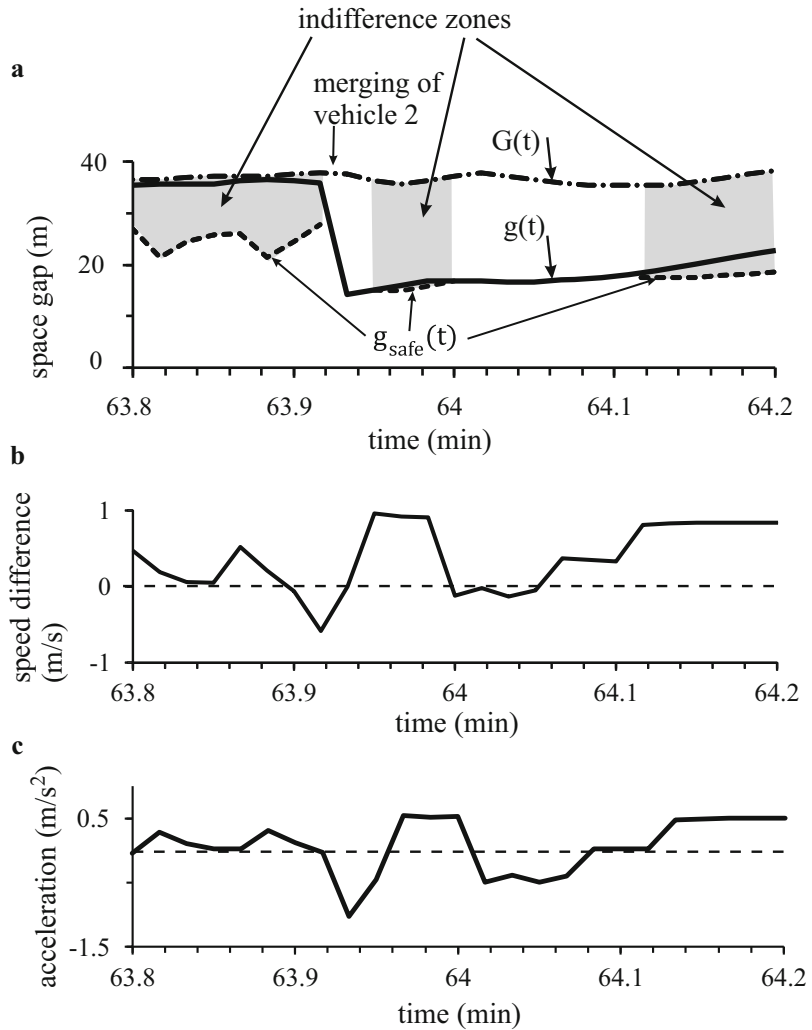
Effect of Platoons of Autonomous Driving Vehicles on Probability of Traffic Breakdown in Mixed Traffic Flow

If the share of autonomous driving vehicles in mixed traffic flow increases, (Fig. 13), the probability of traffic breakdown caused by ACC-vehicles that deteriorate traffic can increase considerably (curve 3 in Fig. 13).

Contrarily to classical ACC-vehicles, long enough platoons of TPACC-vehicles in mixed traffic flow decrease the breakdown probability (curve 2 in Fig. 13). This is explained by small

Autonomous Driving in the Framework of Three-Phase Traffic Theory,

Fig. 12 Continuation of Fig. 10c, d. **(a)** Time function of the space gap $g(t)$ of TPACC-vehicle together with indifference zones (gray regions); $G(t)$ is synchronization space gap; $g_{\text{safe}}(t)$ (dashed curves) is safe space gap; gray regions show time intervals in which condition (12) for indifference zone is satisfied. **(b)** Relative speed $\Delta v = v_f - v$ between the speed of the preceding vehicle and the speed of TPACC-vehicle. **(c)** Acceleration (deceleration) of TPACC-vehicle given in a larger scale as that in Fig. 10d



local speed disturbances caused by TPACC-vehicles in mixed free flow at the bottleneck in comparison with large local speed disturbances caused by classical ACC-vehicles, as already discussed in sections “[Speed Disturbances Occurring by Passing of ACC- and TPACC-Vehicles Through On-Ramp Bottleneck](#),” “[Effect of Single Autonomous Driving Vehicle on Probability of Traffic Breakdown](#),” and “[Explanation of Effect of Single Autonomous Driving Vehicle on Speed Disturbance at Bottleneck](#).”

In particular, the reduction of the probability of traffic breakdown in mixed traffic flow at the bottleneck through already small platoons of

TPACC-vehicles (curve 2 in Fig. 13) is explained by the speed adaptation effect of the three-phase theory that is the basis of TPACC (9): At each vehicle speed, the TPACC-vehicle makes an arbitrary choice in time headway that satisfies conditions (13). In other words, the TPACC-vehicle accepts different values of time headway at different times and does not control a fixed time headway to the preceding vehicle. This dynamic behavior of TPACC-vehicles decreases the amplitude of local speed disturbances at the bottleneck (Fig. 6c, d). This explains why, in contrast to classical ACC-vehicles, TPACC-vehicles decrease the probability of traffic breakdown in mixed traffic flow.

We note that in mixed traffic flow with $\gamma = 20\%$ of classical ACC-vehicles at the on-ramp inflow rate $q_{\text{on}} = 320$ vehicles/h, the probability of traffic breakdown reaches the maximum value $P^{(B)} = 1$ already at $q_{\text{in}} = 1830$ vehicles/h (curve 3 in Fig. 13). However, it should be noted that at parameters of ACC-vehicles used for simulations of curve 3 in Fig. 13, any platoon of the ACC-vehicles satisfies condition (5) for string stability. This means that if, rather than mixed traffic flow (curve 3 in Fig. 13), we consider free flow consisting of $\gamma = 100\%$ of classical ACC-vehicles, then in accordance with results shown in Fig. 5a, no traffic breakdown at $q_{\text{on}} = 320$ vehicles/h and $q_{\text{in}} = 1830$ vehicles/h is realized at the bottleneck. The latter result remains true, even when the flow rate upstream of the bottleneck increases to $q_{\text{in}} = 2000$ vehicles/h.

In other words, as found in (Kerner 2016), if the percentage of the ACC-vehicles in mixed traffic flow increases continuously above the value $\gamma = 20\%$ used in Fig. 13 (curve 3), then there should be some critical percentage of the ACC-vehicles in mixed traffic flow $\gamma_{\text{cr, increase}}$: At $\gamma = \gamma_{\text{cr, increase}}$, the shift of the function $P^{(B)}(q_{\text{sum}})$ to the left in the flow rate axis reaches its maximum. When the percentage of ACC-vehicles increases subsequently, i.e., $\gamma > \gamma_{\text{cr, increase}}$, then the function $P^{(B)}(q_{\text{sum}})$ shifts to the right in the flow rate axis in comparison with the case $\gamma = \gamma_{\text{cr, increase}}$.

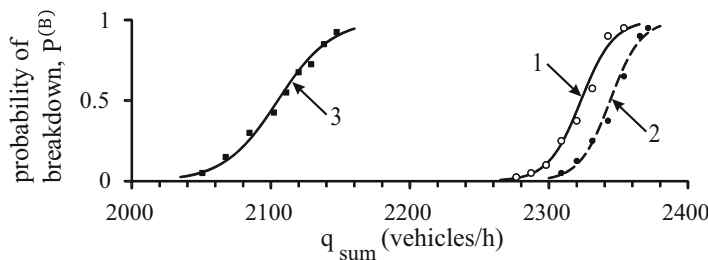
This behavior of the function $P^{(B)}(q_{\text{sum}})$ in mixed traffic flow can be explained as follows (Kerner 2016). At ACC parameters under

consideration (Fig. 13), any platoon of ACC-vehicle is stable. Therefore, when long enough platoons of stable ACC-vehicles begin to build in mixed traffic flow, these platoons can suppress growing speed disturbances in free flow at the bottleneck. We should emphasize that the value $\gamma_{\text{cr, increase}}$ is a large one (in simulations presented in (Kerner 2016) it is $\gamma_{\text{cr, increase}} \approx 35\%$). Therefore, it is related to a nonrealistic case (at least in the near future). In this nonrealistic case, traffic breakdown in mixed traffic flow is mostly determined by behavior of platoons of ACC-vehicles, rather than dynamic interactions between human driving vehicles and ACC-vehicles discussed above (curve 3 in Fig. 13).

Traffic Stream Characteristics of Mixed Traffic Flow

In traffic engineering, the flow–density (the fundamental diagram) and speed–flow relationships are often used to study the effect of traffic control and management on macroscopic traffic stream characteristics (Elefteriadou 2014; Gartner et al. 1997, 2001; Highway Capacity Manual 2000, 2010; May 1990).

To answer a question of how traffic flow is affected when the TPACC strategy versus the ACC strategy is considered (Kerner 2018b), in this section we make a study of traffic stream flow characteristics related to simulations of mixed traffic flow presented in Figs. 5, 6, 7, 8, 9,



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 13 Probability of traffic breakdown at on-ramp bottleneck as a function of the flow rate $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$ at a given on-ramp inflow rate $q_{\text{on}} = 320$ vehicles/h in mixed traffic flow with 20% autonomous driving vehicles: Curve 1 is related to traffic

flow without autonomous driving vehicles. Curves 2 and 3 are related to mixed traffic flow with TPACC- vehicles (curve 2) and ACC-vehicles (curve 3). Simulation parameters of ACC and TPACC are, respectively, the same as those in Figs. 5 and 6, respectively

10, 11, 12, and 13. For simplicity, we consider below macroscopic traffic stream characteristics with the use of speed–flow relationships. The associated flow–density relationships (the fundamental diagram) can be found in (Kerner 2018b).

Traffic Stream Flow Characteristics of Mixed Traffic Flow with 2% Autonomous Driving Vehicles

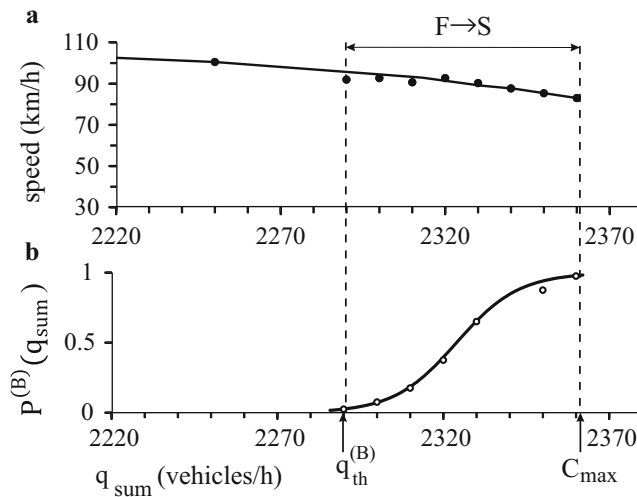
In Fig. 14a, we show a part of the speed–flow relationship for larger flow rates in free flow without autonomous driving vehicles. It turns out that traffic stream flow characteristics are identical for traffic without autonomous driving vehicles and for mixed traffic with 2% of TPACC-vehicles: Single TPACC-vehicles do not affect the stream flow characteristics in free flow (Fig. 14a). This result has already been mentioned in section “Effect of Single Autonomous Driving Vehicle on Probability of Traffic Breakdown,” when we

have discussed curve 1 shown in Fig. 7 for the flow rate dependence of the probability of traffic breakdown $P^{(B)}(q_{\text{sum}})$.

In the three-phase theory (see books (Kerner 2004, 2009, 2017a) and this Encyclopedia (Kerner 2017b, d)), there is a deep connection between the flow rate dependence of the probability of traffic breakdown $P^{(B)}(q_{\text{sum}})$ (Fig. 7) and the overall flow as well as other traffic stream flow characteristics (Fig. 14). In particular, on traffic stream flow characteristics, one should distinguish a flow rate range (Fig. 14)

$$q_{\text{th}}^{(B)} \leq q_{\text{sum}} \leq C_{\text{max}}. \quad (14)$$

Within the flow rate range (1), free flow is in a metastable state with respect to traffic breakdown ($F \rightarrow S$ transition) at the bottleneck (Figs. 1 and 15). A characteristic flow rate $q_{\text{sum}} = q_{\text{th}}^{(B)}$ in (14) is a



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 14 Stream flow characteristics of free flow on single-lane road with on-ramp bottleneck in traffic without autonomous driving vehicles and in mixed traffic with 2% of TPACC-vehicles: (a) A part of speed–flow relationship for larger flow rates. (b) The breakdown probability $P^{(B)}(q_{\text{sum}})$; function $P^{(B)}(q_{\text{sum}})$ is curve 1 from Fig. 7. Traffic stream flow characteristics have been calculated as follows. At each given flow rate q_{sum} (black points on the characteristics), 5-min averaged data for the speed, density, and flow rate have been measured with the use of a virtual road detector installed at the end of the on-ramp merging region $x = 10.3$ km. The data

has been measured *only* during time interval within which free flow has been observed in a simulation realization. Then, as by the calculation of $P^{(B)}(q_{\text{sum}})$ in Fig. 7, $N_r = 40$ different realizations have been simulated for each of the chosen flow rates q_{sum} . This allows us to make a statistical analysis of the average speed and density in the traffic stream. Black points on the speed–flow relationship are related to the average values of the speed and density derived from this statistical analysis. Other model parameters are the same as those in Fig. 6. Calculated values: $q_{\text{th, TPACC}}^{(B)} = q_{\text{th}}^{(B)} = 2290$ and $C_{\text{max, TPACC}} = C_{\text{max}} = 2360$ vehicles/h

threshold flow rate for spontaneous traffic breakdown at the bottleneck: At $q_{\text{sum}} < q_{\text{th}}^{(B)}$ the breakdown probability $P^{(B)} = 0$ (Fig. 15a), i.e., no spontaneous traffic breakdown can occur during a time interval of the observation of traffic flow T_{ob} .

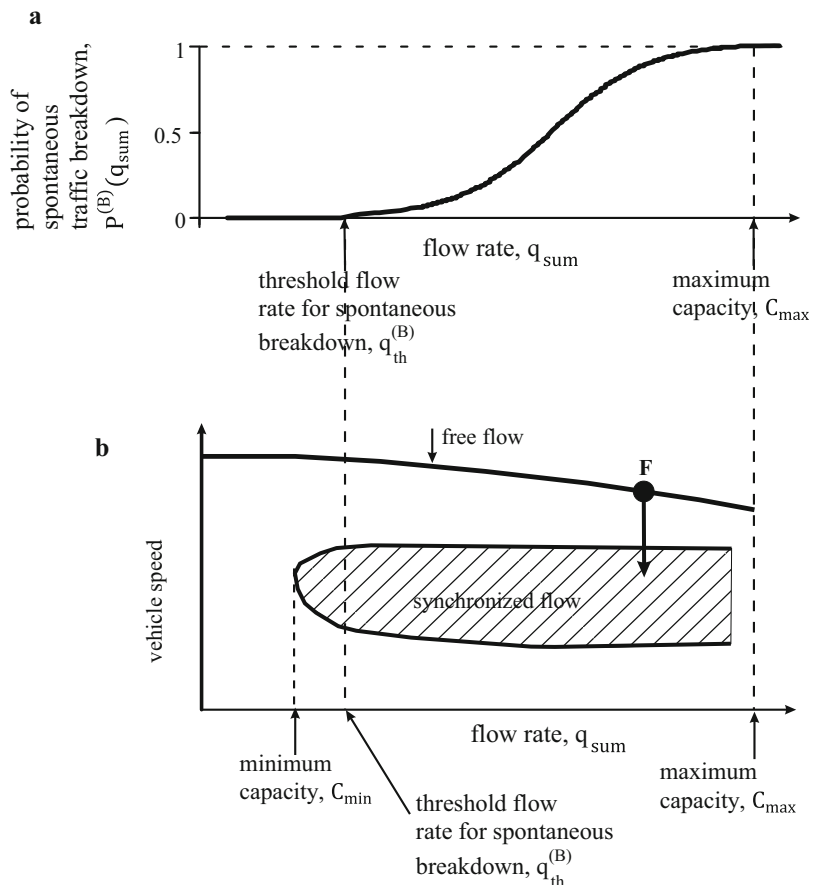
A characteristic flow rate $q_{\text{sum}} = C_{\text{max}}$ in (1) and (14) is the maximum highway capacity (Figs. 1 and 15): At $q_{\text{sum}} \geq C_{\text{max}}$ the breakdown probability $P^{(B)} = 1$ (Fig. 15a), i.e., spontaneous traffic breakdown does occur at the bottleneck during the time interval T_{ob} .

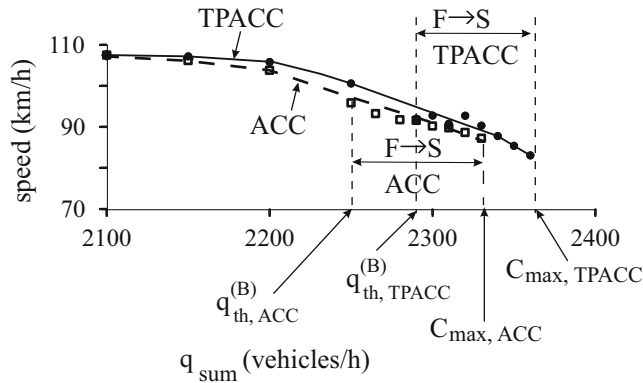
The larger the values $q_{\text{th}}^{(B)}$ and C_{max} for the traffic stream, the larger is on average the overall flow. Therefore, the characteristic flow rates $q_{\text{th}}^{(B)}$ and C_{max} , which determine the boundaries of the flow rate range (14) (Fig. 15), are basic statistical characteristics of the overall flow in the traffic stream in the framework of the three-phase theory. A more detailed discussion of the application of the three-

phase theory for a statistical analysis of traffic stream characteristics (like the definition of “stochastic highway capacity of free flow at a bottleneck” and its theoretical justification made in the three-phase theory) as well as a critical consideration of the three-phase theory (Kerner 1999a, b, c, 2004, 2009, 2013, 2017a) versus the classical traffic flow theories and models reviewed in (Bellomo et al. 2002; Brockfeld et al. 2003; Chowdhury et al. 2000; Daganzo 1997; Elefteriadou 2014; Ferrara et al. 2018; Gartner et al. 1997, 2001; Gazis 2002; Haight 1963; Helbing 2001; Leutzbach 1988; Mahnke et al. 2005, 2009; Nagatani 2002, Nagel et al. 2003; Newell 1982; Papageorgiou 1983; Saifuzzaman and Zheng 2014; Schadschneider et al. 2011; Treiber and Kesting 2013; Whitham 1974; Wiedemann 1974) are out of scope of this entry: Such a detailed analysis has already been

Autonomous Driving in the Framework of Three-Phase Traffic Theory,

Fig. 15 Qualitative explanation of condition (14): (a) Qualitative flow rate function of the breakdown probability $P^{(B)}(q_{\text{sum}})$. (b) Z-characteristic of traffic breakdown with states for free flow and synchronized flow taken from Fig. 1. Adapted from (Kerner 2017a)





Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 16 Comparison of traffic stream flow characteristics for free flow on single-lane road with on-ramp bottleneck in mixed traffic with 2% of autonomous driving vehicles. Parts of speed–flow relationship for larger flow rates. Solid curve “TPACC” is related to TPACC-vehicles. Dashed curve “ACC” is related to

classical ACC-vehicles. Stream flow characteristics have been calculated as explained in caption Fig. 14. Other model parameters are the same as those in Figs. 5 and 6, respectively. Calculated values $q_{th, TPACC}^{(B)} = 2290$ and $C_{max, TPACC} = 2360$ vehicles/h; $q_{th, ACC}^{(B)} = 2265$ and $C_{max, ACC} = 2330$ vehicles/h

made in the book (Kerner 2017a) as well as in this Encyclopedia (Kerner 2017d).

For the further analysis, we denote the flow rate range (14) on traffic stream characteristics by the arrow “F→S” (Figs. 14, 16, and 17).

We denote the statistical characteristics of the overall flow $q_{th}^{(B)}$, C_{max} in (14) for mixed traffic flow by $q_{th, TPACC}^{(B)}$, $C_{max, TPACC}$, when autonomous driving vehicles are TPACC-vehicles, and by $q_{th, ACC}^{(B)}$, $C_{max, ACC}$ for classical ACC-vehicles, respectively. We have found that the overall flow characteristics do not change on average in mixed traffic with 2% of TPACC-vehicles (Fig. 14): $q_{th, TPACC}^{(B)} = q_{th}^{(B)}$ and $C_{max, TPACC} = C_{max}$.

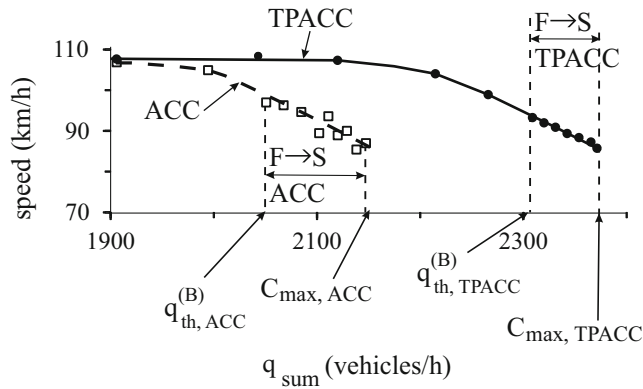
In contrast with mixed traffic flow with 2% of TPACC-vehicles, we have found that both values $q_{th, ACC}^{(B)}$ and $C_{max, ACC}$ decrease in mixed traffic flow with 2% of classical ACC-vehicles (Fig. 16, curve “ACC”). This means that already 2% of classical ACC-vehicles reduce on average the overall flow in the traffic stream. As explained in sections “Speed Disturbances Occurring by Passing of ACC- and TPACC-Vehicles Through On-Ramp Bottleneck” and “Effect of Single Autonomous Driving Vehicle on Probability of Traffic Breakdown,” this result is associated with a large local speed disturbance caused by a

classical ACC- vehicle at the bottleneck: Within the flow rate range (14), the large local speed disturbance can initiate a nucleus for spontaneous traffic breakdown at the bottleneck.

Traffic Stream Flow Characteristics of Mixed Traffic Flow with 20% Autonomous Driving Vehicles

We might assume that vehicles implementing TPACC strategy (9) can reduce the overall flow for the traffic stream due to their different use of available space on the road. However, simulations presented in Figs. 4, 5, 6, 7, 8, 9, 10, 11, 12, and 13 show that no such adverse effect for the traffic stream occurs. However, rather than reduction of the overall flow through the use of TPACC- vehicles, we have found that TPACC-vehicles increase on average the overall flow (compare values $q_{th, TPACC}^{(B)}$, $C_{max, TPACC}$ in Fig. 14 with, respectively, these values given in caption to Fig. 17).

In contrast with TPACC-vehicles, we have found that classical ACC-vehicles reduce on average the overall flow in mixed traffic flow (Fig. 17). As explained in sections “Speed Disturbances Occurring by Passing of ACC- and TPACC-Vehicles Through On-Ramp Bottleneck” and “Effect of Single Autonomous Driving Vehicle on Probability of



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 17 Comparison of traffic stream flow characteristics for free flow on single-lane road with on-ramp bottleneck in mixed traffic with 20% of autonomous driving vehicles. Parts of speed-flow relationships for larger flow rates. Solid curve “TPACC” is related to TPACC-vehicles. Dashed curve “ACC” is related to classical ACC-vehicles. Stream flow characteristics have been calculated as explained in caption to

Fig. 14. In simulations, at $q_{\text{sum}} \leq 2000$ vehicles/h, there is no on-ramp inflow ($q_{\text{on}} = 0$); at $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}} > 2000$ vehicles/h, the increase in q_{sum} has been achieved through increase in q_{on} at constant $q_{\text{in}} = 2000$ vehicles/h. Other model parameters are the same as those in Figs. 5 and 6, respectively. Calculated values: $q_{\text{th, TPACC}}^{(B)} = 2308$ and $C_{\text{max, TPACC}} = 2371$ vehicles/h; $q_{\text{th, ACC}}^{(B)} = 2050$ and $C_{\text{max, ACC}} = 2147$ vehicles/h

Traffic Breakdown,” this effect of the overall flow reduction caused by classical ACC-vehicles is explained by the occurrence of large local speed disturbances at the bottleneck in mixed traffic flow. The local speed disturbances initiate traffic breakdown in the mixed traffic flow at considerably smaller flow rates in comparison with traffic flow without classical ACC-vehicles. Thus, through the strong effect of autonomous driving vehicles on traffic breakdown at the bottleneck, in simulations we cannot resolve the effect of the different use of available space on the road by TPACC-vehicles and ACC-vehicles on the overall flow.

Simulations show that even in a nonrealistic case of traffic consisting of 100% of automated driving vehicles, the effect of large local speed disturbances at the bottleneck on the overall flow caused by the classical ACC (Fig. 5) is also stronger than the effect of the different use of available space on the road by the ACC-vehicles and TPACC-vehicles. Indeed, under simulation parameters of ACC and TPACC used in Figs. 5 and 6, the maximum highway capacity for TPACC-vehicles $C_{\text{max, TPACC}} = 2353$ vehicles/h is slightly larger than that for ACC-vehicles $C_{\text{max, ACC}} = 2322.6$ vehicles/h.

Thus, we have found that autonomous driving based on the TPACC strategy can increase on average the overall flow for mixed traffic flow. Contrarily to the TPACC strategy, autonomous driving based on the classical ACC strategy decreases on average the overall flow (Figs. 16 and 17).

Discussion

About Applicability of Model Results for Future Autonomous Driving in Mixed Traffic Flow

We have shown that a TPACC model (9) (Kerner 2017c, 2018b) has allowed us to understand the physics of TPACC-vehicles in mixed traffic flow. In particular, we have found the effect of TPACC-vehicles on the probability of traffic breakdown at a road bottleneck. However, the following question can arise: Whether can the simple TPACC model (9) be applicable for reliable statements about physical features of real mixed traffic flow consisting of human driving vehicles and TPACC-vehicles? To answer this question, we consider firstly some features of the TPACC model (9) that might be seemed at the first glance as nonrealistic for real traffic flow.

It seems that TPACC model (9) is mathematically of small incremental value from the pre-existing classical ACC model (4). However, we have above shown that this “small incremental mathematical value” exhibits a large physical effect on traffic flow. This large physical effect on traffic flow through the TPACC strategy (9) is associated with the TPACC physical feature mentioned above: Through the indifference zone of TPACC (9), a TPACC-vehicle does not react on the time headway change within the time headway range (13). This dynamic behavior of TPACC-vehicles decreases local speed disturbances in free flow at the bottleneck. The reduction of the local speed disturbances results in a decrease in the breakdown probability in the traffic stream.

This review deals with a subset of the functionality required for autonomous driving, namely, longitudinal following a given leader (TPACC). Other challenges for autonomous driving such as lateral dynamics or sensor-related problems, which are important to satisfy a safety motion of autonomous driving vehicles on multilane highways and urban areas (see, e.g., (Bengler et al. 2014; Ioannou 1997; Ioannou and Sun 1996; Ioannou and Kosmatopoulos 2000; Maurer et al. 2015; Meyer and Beiker 2014; Rajamani 2012)), are not tackled in this review entry. Therefore, a question can arise in what degree results derived for TPACC are related to future autonomous driving.

As mentioned in section “Introduction,” in empirical data the qualitative flow rate dependence of the probability of traffic breakdown at a road bottleneck does not depend on the number of highway lanes (on features of lateral dynamics of vehicles), on the bottleneck type, and on real vehicle technology (during last 30 years vehicle technology was changed considerably; however, qualitative empirical features of traffic breakdown did not change). In accordance with the three-phase theory that explains all known empirical features of traffic breakdown (Kerner 2004, 2009, 2017a), the simple ACC model and TPACC model used in the entry reflect dynamic vehicle features that are responsible for traffic breakdown. For this reason, the result of the entry is that at the same model parameters classic

ACC-vehicles (4) increase the breakdown probability, whereas TPACC-vehicles (9) decrease the breakdown probability which proves that the use of indifference zones of the three-phase theory can have benefits for future autonomous driving. This is because contrarily with (4), human driving vehicles do not control time headway within the time headway range (13) (Kerner 2004, 2009, 2017a). Thus, the TPACC-vehicles, which can be considered autonomous driving “learning” from empirical human driving behavior, can decrease the breakdown probability.

Results of this review allow us to assume that future systems for autonomous driving should be developed whose rules are consistent with those of human driving vehicles. Otherwise, we could expect that autonomous driving vehicles can be considered as “obstacles” for drivers. The physics of autonomous driving in the framework of the three-phase theory studied in this entry emphasizes that future autonomous driving should be developed in which both the longitudinal dynamics (TPACC) and lateral dynamics should learn from driver behavior. In particular, the longitudinal and lateral dynamics of autonomous driving vehicles should be consistent with the existence of indifference zones of the three-phase theory (Kerner 1999a, c, 2004, 2009, 2017a).

Conclusions

In the review entry, based on numerical simulations of a simple model for autonomous driving in the framework of the three-phase theory (TPACC) introduced in the entry, we have found that applications of the TPACC strategy can lead to the following advantages in comparison with the classical approach to ACC:

- (i) The absence of string instability.
- (ii) Considerably smaller speed disturbances at road bottlenecks.
- (iii) TPACC-vehicles can decrease the probability of traffic breakdown at the bottleneck in mixed traffic flow; on the contrary, even a single autonomous driving vehicle based on the classical approach can provoke traffic breakdown at the bottleneck in mixed traffic flow.

These advantages of TPACC are associated with the absence of a fixed desired time headway to the preceding vehicle in the TPACC strategy: A TPACC-vehicle exhibits a large indifference zone within the time headway range (13) within which the TPACC-vehicle does not control time headway to the preceding vehicle. As we have found in this entry, due to the large indifference zone within the time headway range (13), the TPACC-vehicle should not necessarily decelerate as strong as the preceding vehicle when a local short-time speed disturbance appears at a road bottleneck. This dynamic behavior of TPACC-vehicles decreases local speed disturbances in free flow at a road bottleneck. In its turn, the decrease in the amplitude of local speed disturbances at road bottlenecks results in a decrease in the probability of traffic breakdown in the traffic stream.

Future Directions

In this review entry, we have made a comparison of the effect of classical ACC-vehicles and TPACC-vehicles on the probability of traffic breakdown at a road bottleneck in mixed traffic flow. In this scenario of the application of autonomous driving vehicles, we have considered only some specific values of TPACC and ACC parameters (e.g., K_1 , K_2 , and headway time) to demonstrate that the TPACC strategy can exhibit advantages in comparison with the classical ACC.

However, larger values of the dynamic coefficients K_1 and K_2 of ACC (4) and shorter headway time might give entirely different results. Moreover, the examples presented in the entry only involve situations where the total flow is near capacity. Thus, without exploring a wide range of scenarios, we cannot make a general claim that the TPACC system is superior to ACC systems. Additionally, incorporating cooperative merging between ACC-vehicles could reduce the tendency to initiate breakdown at highway bottlenecks. We believe that related detailed studies of the TPACC model introduced in (Kerner 2017c, 2018b), which are out of the scope of this review, will be a very interesting task of future investigations of the physics of autonomous driving.

We can expect that other objectives of a comparison of the TPACC strategy versus the ACC strategy can be interesting for further studies. Examples can be a study of (i) congested mixed traffic build at the bottleneck after traffic breakdown has occurred or (ii) characteristics of mixed traffic flow consisting of autonomous driving and human driving vehicles with different driver and vehicle parameters or (iii) the effect of the TPACC strategy on multilane mixed traffic with different types of bottlenecks.

Acknowledgments I would like to thank Sergey Klenov for the help and useful suggestions. We thank our partners for their support in the project “MEC-View – Object detection for autonomous driving based on Mobile Edge Computing,” funded by the German Federal Ministry of Economic Affairs and Energy.

Appendix A. Kerner–Klenov Microscopic Stochastic Traffic Flow Model

In this Appendix, we make explanations of the Kerner–Klenov stochastic microscopic three-phase model for human driving vehicles (Kerner and Klenov 2002, 2003, 2009) and model parameters used for simulations of mixed traffic flow presented in the main text.

Update Rules of Vehicle Motion

In a discrete model version of the Kerner–Klenov stochastic microscopic three-phase model used in all simulations presented in the main text, rather than the continuum space coordinate (Kerner and Klenov 2002), a discretized space coordinate with a small enough value of the discretization space interval δx is used (Kerner and Klenov 2009). Consequently, the vehicle speed and acceleration (deceleration) discretization intervals are $\delta v = \delta x/\tau$ and $\delta a = \delta v/\tau$, respectively, where τ is time step. Because in the discrete model version discrete (and dimensionless) values of space coordinate, speed, and acceleration are used, which are measured, respectively, in values δx , δv , and δa , and time is measured in values of τ , value τ in all formulas is assumed below to be the dimensionless value $\tau = 1$. In the

discrete model version used for all simulations, the discretization cell $\delta x = 0.01$ m is used.

A choice of $\delta x = 0.01$ m made in the model determines the accuracy of vehicle speed calculations *in comparison* with the initial continuum in space stochastic model of (Kerner and Klenov 2002). We have found that the discrete model exhibits similar characteristics of phase transitions and resulting congested patterns at highway bottlenecks as those in the continuum model at δx that satisfies the conditions

$$\delta x / \tau^2 \ll b, a, a^{(a)}, a^{(b)}, a^{(0)}, \quad (15)$$

where model parameters for driver deceleration and acceleration $b, a, a^{(a)}, a^{(b)}$, and $a^{(0)}$ will be explained below.

Update rules of vehicle motion in the discrete model for identical drivers and identical vehicles moving in a road lane are as follows (Kerner and Klenov 2009):

$$v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}_{n+1} + \xi_n, v_n + a\tau, v_{s,n})), \quad (16)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (17)$$

where the index n corresponds to the discrete time $t_n = \tau n$, $n = 0, 1, \dots$, v_n is the vehicle speed at time step n , a is the maximum acceleration, and $\tilde{v}_n$ is the vehicle speed without speed fluctuations ξ_n :

$$\tilde{v}_{n+1} = \min(v_{\text{free}}, v_{s,n}, v_{c,n}), \quad (18)$$

$$v_{c,n} = \begin{cases} v_n + \Delta_n & \text{at } g_n \leq G_n \\ v_n + a_n\tau & \text{at } g_n > G_n, \end{cases} \quad (19)$$

$$\Delta_n = \max(-b_n\tau, \min(a_n\tau, v_{\ell,n} - v_n)), \quad (20)$$

$$g_n = x_{\ell,n} - x_n - d. \quad (21)$$

The subscript ℓ marks variables related to the preceding vehicle, $v_{s,n}$ is a safe speed at time step n , v_{free} is the free flow speed in free flow, ξ_n describes speed fluctuations, g_n is a space gap between two vehicles following each other, and

G_n is the synchronization space gap; all vehicles have the same length d . The vehicle length d includes the mean space gap between vehicles that are in a standstill within a wide moving jam. Values $a_n \geq 0$ and $b_n \geq 0$ in (19) and (20) restrict changes in speed per time step when the vehicle accelerates or adjusts the speed to that of the preceding vehicle.

Synchronization Space Gap and Hypothetical Steady States of Synchronized Flow

Equations (19) and (20) describe the adaptation of the vehicle speed to the speed of the preceding vehicle, i.e., the speed adaptation effect in synchronized flow. This vehicle speed adaptation takes place within the synchronization gap G_n : At

$$g_n \leq G_n \quad (22)$$

the vehicle tends to adjust its speed to the speed of the preceding vehicle. This means that the vehicle decelerates if $v_n > v_{\ell,n}$ and accelerates if $v_n < v_{\ell,n}$.

In (19), the synchronization gap G_n depends on the vehicle speed v_n and on the speed of the preceding vehicle $v_{\ell,n}$:

$$G_n = G(v_n, v_{\ell,n}), \quad (23)$$

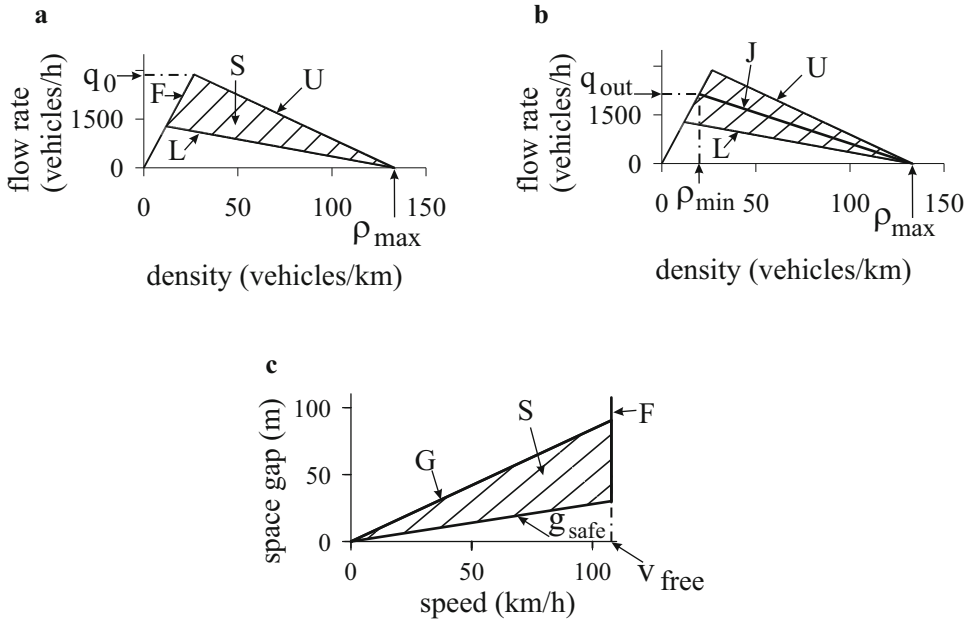
$$G(u, w) = \max(0, \lfloor k\tau u + a^{-1}u(u - w) \rfloor), \quad (24)$$

where $k > 1$ is constant; $\lfloor z \rfloor$ denotes the integer part of z .

The speed adaptation effect within the synchronization distance is related to the hypothesis of the three-phase theory: Hypothetical steady states of synchronized flow cover a 2D region in the flow-density (Fig. 18a). Boundaries F , L , and U of this 2D region shown in Fig. 18a are, respectively, associated with the free flow speed in free flow, a synchronization space gap G , and a safe space gap g_{safe} . A speed function of the safe space gap $g_{\text{safe}}(v)$ is found from the equation

$$v = v_s(g_{\text{safe}}, v). \quad (25)$$

Respectively, as for the continuum model (see Sec. 16.3 of the book (Kerner 2004)), for the



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 18 Steady speed states for the Kerner-Klenov traffic flow model in the flow-density (a, b) and in the space gap-speed planes (c). In (a, b), L and U are, respectively, lower and upper boundaries of 2D regions of steady states of synchronized flow. In (b), J is

the line J whose slope is equal to the characteristic mean velocity v_g of the downstream front of a wide moving jam; in the flow-density plane, the line J represents the propagation of the downstream front of the wide moving jam with time-independent velocity v_g . F, free flow; S, synchronized flow

discrete model hypothetical steady states of synchronized flow cover a 2D region in the flow-density plane (Fig. 18a, b). However, because the speed v and space gap g are integers in the discrete model, the steady states do not form a continuum in the flow-density plane as they do in the continuum model. The inequalities

$$v \leq v_{\text{free}}, \quad g \leq G(v, v), \quad g \geq g_{\text{safe}}(v) \quad (26)$$

define a 2D region in the space gap-speed plane (Fig. 18c) in which the hypothetical steady states exist for the discrete model, when all model fluctuations are neglected.

In (26), we have taken into account that in the hypothetical steady states of synchronized flow vehicle speeds and space gaps are assumed to be time-independent, and the speed of each of the vehicles is equal to the speed of the associated preceding vehicle: $v = v_{\ell}$. However, due to model fluctuations, steady states of synchronized flow are destroyed, i.e., they do not exist in simulations; this

explains the term “hypothetical” steady states of synchronized flow. Therefore, rather than steady states, some nonhomogeneous in space and time traffic states occur. In other words, steady states are related to a hypothetical model fluctuation-less limit of homogeneous in space and time vehicle motion that does not realized in real simulations. Driver time delays are described through model fluctuations. Therefore, any application of the Kerner-Klenov stochastic microscopic three-phase traffic flow model without model fluctuations has *no sense*. In other words, for the description of real spatiotemporal traffic flow phenomena, model speed fluctuations incorporated in this model are needed.

Model Speed Fluctuations

In the model, random vehicle deceleration and acceleration are applied depending on whether the vehicle decelerates or accelerates or else maintains its speed:

$$\zeta_n = \begin{cases} \zeta_a & \text{if } S_{n+1} = 1 \\ -\zeta_b & \text{if } S_{n+1} = -1 \\ \zeta^{(0)} & \text{if } S_{n+1} = 0. \end{cases} \quad (27)$$

State of vehicle motion S_{n+1} in (27) is determined by formula

$$S_{n+1} = \begin{cases} -1 & \text{if } \tilde{v}_{n+1} < v_n \\ 1 & \text{if } \tilde{v}_{n+1} > v_n \\ 0 & \text{if } \tilde{v}_{n+1} = v_n. \end{cases} \quad (28)$$

In (27), ζ_b , $\zeta^{(0)}$, and ζ_a are random sources for deceleration and acceleration that are as follows:

$$\zeta_b = a^{(b)} \tau \Theta(p_b - r), \quad (29)$$

$$\zeta^{(0)} = a^{(0)} \tau \begin{cases} -1 & \text{if } r < p^{(0)} \\ 1 & \text{if } p^{(0)} \leq r < 2p^{(0)} \text{ and } v_n > 0 \\ 0 & \text{otherwise,} \end{cases} \quad (30)$$

$$\zeta_a = a^{(a)} \tau \Theta(p_a - r), \quad (31)$$

p_b is the probability of random vehicle deceleration, p_a is the probability of random vehicle acceleration, $p^{(0)}$ and $a^{(0)} \leq a$ are constants, $r = \text{rand}(0, 1)$, $\Theta(z) = 0$ at $z < 0$ and $\Theta(z) = 1$ at $z \geq 0$, and $a^{(a)}$ and $a^{(b)}$ are model parameters (see Table 1), which in some applications can be chosen as speed functions $a^{(a)} = a^{(a)}(v_n)$ and $a^{(b)} = a^{(b)}(v_n)$.

Stochastic Time Delays of Acceleration and Deceleration

To simulate time delays either in vehicle acceleration or in vehicle deceleration, a_n and b_n in (20) are taken as the following stochastic functions

$$a_n = a \Theta(P_0 - r_1), \quad (32)$$

$$b_n = a \Theta(P_1 - r_1), \quad (33)$$

$$P_0 = \begin{cases} p_0 & \text{if } S_n \neq 1 \\ 1 & \text{if } S_n = 1, \end{cases} \quad (34)$$

$$P_1 = \begin{cases} p_1 & \text{if } S_n \neq -1 \\ p_2 & \text{if } S_n = -1, \end{cases} \quad (35)$$

$r_1 = \text{rand}(0, 1)$, p_1 is constant, and $p_0 = p_0(v_n)$ and $p_2 = p_2(v_n)$ are speed functions (see Table 1).

Autonomous Driving in the Framework of Three-Phase Traffic Theory, Table 1 Model parameters of vehicle motion in road lane used in simulations of the main text

$\tau_{\text{safe}} = \tau = 1 \text{ s}, d = 7.5 \text{ m}/\delta x,$
$\delta x = 0.01 \text{ m}, \delta v = 0.01 \text{ ms}^{-1}, \delta a = 0.01 \text{ ms}^{-2},$
$v_{\text{free}} = 30 \text{ ms}^{-1}/\delta v, b = 1 \text{ ms}^{-2}/\delta a, a = 0.5 \text{ ms}^{-2}/\delta a,$
$k = 3, p_1 = 0.3, p_b = 0.1, p_a = 0.17, p^{(0)} = 0.005,$
$p_0(v_n) = 0.575 + 0.125 \min(1, v_n/v_{01}),$
$p_2(v_n) = 0.48 + 0.32 \Theta(v_n - v_{21}),$
$v_{01} = 10 \text{ ms}^{-1}/\delta v, v_{21} = 15 \text{ ms}^{-1}/\delta v,$
$a^{(0)} = 0.2a, a^{(a)} = a^{(b)} = a.$

Simulations of Slow-To-Start Rule

In the model, simulations of the well-known effect of the driver time delay in acceleration at the downstream front of synchronized flow or a wide moving jam known as a slow-to-start rule (Barlović et al. 1998; Takayasu and Takayasu 1993) are made as a collective effect through the use of Eqs. (19) and (20) and a random value of vehicle acceleration (32). Eq. (32) with $P_0 = p_0 < 1$ is applied only if the vehicle did not accelerate at the former time step ($S_n \neq 1$); in the latter case, a vehicle accelerates with some probability p_0 that depends on the speed v_n ; otherwise $P_0 = 1$ (see formula (34)).

The mean time delay in vehicle acceleration is equal to

$$\tau_{\text{del}}^{(\text{acc})}(v_n) = \frac{\tau}{p_0(v_n)}. \quad (36)$$

From formula (36), it follows that the mean time delay in vehicle acceleration from a standstill within a wide moving jam (i.e., when in formula (36) the speed $v_n = 0$) is equal to

$$\tau_{\text{del}}^{(\text{acc})}(0) = \frac{\tau}{p_0(0)}. \quad (37)$$

The mean time delay in vehicle acceleration from a standstill within a wide moving jam determines the parameters of the line J in the flow-density plane (Fig. 18b).

Probability $p_0(v_n)$ in (34) is chosen to be an increasing speed function (see Table 1). Because

the speed within synchronized flow is larger than zero, the mean time delay in vehicle acceleration at the downstream front of synchronized flow that we denote by

$$\tau_{\text{del}, \text{syn}}^{(\text{acc})} = \tau_{\text{del}}^{(\text{acc})}(v_n), \quad v_n > 0 \quad (38)$$

is shorter than the mean time delay in vehicle acceleration at the downstream front of the wide moving jam $\tau_{\text{del}}^{(\text{acc})}(0)$:

$$\tau_{\text{del}, \text{syn}}^{(\text{acc})} < \tau_{\text{del}}^{(\text{acc})}(0). \quad (39)$$

Safe Speed

In the model, the safe speed $v_{s,n}$ in (16) is chosen in the form

$$v_{s,n} = \min\left(v_n^{(\text{safe})}, g_n/\tau + v_\ell^{(a)}\right), \quad (40)$$

$v_\ell^{(a)}$ is an “anticipation” speed of the preceding vehicle that will be considered below, and the function

$$v_n^{(\text{safe})} = \left\lfloor v^{(\text{safe})}(g_n, v_{\ell,n}) \right\rfloor \quad (41)$$

in (40) is related to the safe speed $v^{(\text{safe})}(g_n, v_{\ell,n})$ in the model by Krauß et al. (1997), which is a solution of the Gipps’s Eq. [32]

$$v^{(\text{safe})}\tau + X_d\left(v^{(\text{safe})}\right) = g_n + X_d(v_{\ell,n}), \quad (42)$$

where $X_d(u)$ is the braking distance that should be passed by the vehicle moving first with the speed u before the vehicle can come to a stop.

The condition (42) enables us to find the safe speed $v^{(\text{safe})}$ as a function of the space gap g_n and speed $v_{\ell,n}$ provided $X_d(u)$ is a known function. In the case when the vehicle brakes with a constant deceleration b , the change in the vehicle speed for each time step is $-b\tau$ except the last time step before the vehicle comes to a stop. At the last time step, the vehicle decreases its speed at the value $b\tau\beta$, where β is a fractional part of $u/b\tau$. According to formula (17) for the displacement of the vehicle for one time step, the braking distance $X_d(u)$ is (Krauß et al. 1997)

$$X_d(u) = \tau(u - b\tau + u - 2b\tau + \dots + \beta b\tau). \quad (43)$$

From (43), it follows (Krauß et al. 1997)

$$X_d(u) = b\tau^2 \left(\alpha\beta + \frac{\alpha(\alpha-1)}{2} \right), \quad (44)$$

$\alpha = \lfloor u/b\tau \rfloor$ is an integer part of $u/b\tau$.

The safe speed $v^{(\text{safe})}$ as a solution of Eq. (42) at the distance $X_d(u)$ given by (44) has been found by Krauß et al. (1997)

$$v^{(\text{safe})}(g_n, v_{\ell,n}) = b\tau(\alpha_{\text{safe}} + \beta_{\text{safe}}), \quad (45)$$

where

$$\alpha_{\text{safe}} = \left\lfloor \sqrt{2 \frac{X_d(v_{\ell,n}) + g_n}{b\tau^2} + \frac{1}{4} - \frac{1}{2}} \right\rfloor, \quad (46)$$

$$\beta_{\text{safe}} = \frac{X_d(v_{\ell,n}) + g_n}{(\alpha_{\text{safe}} + 1)b\tau^2} - \frac{\alpha_{\text{safe}}}{2}. \quad (47)$$

The safe speed in the model by Krauß et al. (1997) provides collisionless motion of vehicles if the time gap g_n/v_n between two vehicles is greater than or equal to the time step τ , i.e., if $g_n \geq v_n\tau$ (Krauß 1998). In the model, it is assumed that in some cases, mainly due to lane changing or merging of vehicles onto the main road within the merging region of bottlenecks, the space gap g_n can become less than $v_n\tau$. In these critical situations, the collision-less motion of vehicles in the model is a result of the second term in (40) in which some prediction $(v_\ell^{(a)})$ of the speed of the preceding vehicle at the next time step is used. The related “anticipation” speed $v_\ell^{(a)}$ at the next time step is given by formula

$$v_\ell^{(a)} = \max\left(0, \min\left(v_{\ell,n}^{(\text{safe})}, v_{\ell,n}, g_{\ell,n}/\tau\right) - a\tau\right), \quad (48)$$

where $v_{\ell,n}^{(\text{safe})}$ is the safe speed (41) and (45, 46, 47) for the preceding vehicle and $g_{\ell,n}$ is the space gap in front of the preceding vehicle. Simulations have shown that formulas (40), (41), and (45), (46),

(47), (48) lead to collision-less vehicle motion over a wide range of parameters of the merging region of on-ramp bottlenecks (Appendix D).

In hypothetical steady states of traffic flow (Fig. 18a), the safe space gap g_{safe} is determined from condition $v = v_s$. In accordance with Eqs. (40, 41, and 42), at a given v in steady traffic states $v = v_\ell$ for the safe speed v_s , we get

$$v_s = g_{\text{safe}}/\tau, \quad (49)$$

and, therefore,

$$g_{\text{safe}} = v\tau. \quad (50)$$

Thus, for hypothetical steady states of traffic flow $\tau_{\text{safe}} = \tau = 1$ s.

In (Kerner 2017a) it has been shown that the Kerner–Klenov model (16)–(21), (23), (24), (27)–(35), (40), (41), and (45)–(48) is a Markov chain: At time step $n + 1$, values of model variables v_{n+1} , x_{n+1} , and S_{n+1} are calculated based only on their values v_n , x_n , and S_n at step n .

It should be noted that after the Kerner–Klenov car-following model with indifference zones in car-following based on the three-phase theory (dashed region in Fig. 3) has been introduced (Kerner and Klenov 2002) (as mentioned, in this review entry a discrete version of this model of Ref. (Kerner and Klenov 2009) has been used), a number of different traffic flow models incorporating some of the hypotheses of the three-phase theory have been developed, and many results with the use of these models have been found (e.g., (Borsche et al. 2012; Davis 2004a, b, 2006a, b, 2007, 2008, 2014; Gao et al. 2007, 2009; Hausken and Rehborn 2015; He et al. 2010; Hoogendoorn et al. 2008; Jiang and Wu 2004, 2005, 2007a; Jiang et al. 2007, 2014, 2015, 2017; Jiang and Wu 2007b; Jin et al. 2010, 2015; Jin and Wang 2011; Kerner 2015b; Kerner and Klenov 2003, 2006, 2018; Kerner et al. 2002, 2006a, b, 2007, 2011, 2013a, b, 2014; Klenov 2010; Gasnikov et al. 2013; Kokubo et al. 2011; Knorr and Schreckenberg 2013; Lee et al. 2004; Lee and Kim 2011; Li et al. 2007; Neto et al. 2011; Qian et al. 2017; Pottmeier et al. 2007; Rehborn and Klenov 2009;

Rehborn et al. 2011a, b, 2017; Rehborn and Koller 2014; Siebel and Mauser 2006; Tian et al. 2009, 2012, 2016a, b, c; Wang et al. 2007; Wu et al. 2008; Xiang et al. 2013; Yang et al. 2013, 2018; Zhang et al. 2012)). A review of some of these models that are related to the class of cellular automation models that can be found in this Encyclopedia has been recently done by Tian et al. (2018).

Appendix B. Discrete Version of Classical ACC Model

In simulations of the classical ACC model (4), as in the model of human driving vehicles (Appendix A), we use the discrete time $t = n\tau$, where $n = 0, 1, 2, \dots$; $\tau = 1$ s is time step. Therefore, the space gap to the preceding vehicle is equal to $g_n = x_{\ell,n} - x_n - d$, and the relative speed is given by $\Delta v_n = v_{\ell,n} - v_n$, where x_n and v_n are coordinate and speed of the ACC-vehicle, $x_{\ell,n}$ and $v_{\ell,n}$ are coordinate and speed of the preceding vehicle, and d is the vehicle length that is assumed the same one for autonomous driving and human driving vehicles. Correspondingly, the classical model of the dynamics of ACC-vehicle (4) can be rewritten as follows (Kerner 2016, 2017a):

$$a_n^{(\text{ACC})} = K_1 \left(g_n - v_n \tau_d^{(\text{ACC})} \right) + K_2 (v_{\ell,n} - v_n). \quad (51)$$

The ACC-vehicles move in accordance with Eq. (51) where, in addition, the following formulas are used:

$$v_{c,n}^{(\text{ACC})} = v_n + \tau \max \left(-b_{\text{max}}, \min \left(\left\lfloor a_n^{(\text{ACC})} \right\rfloor, a_{\text{max}} \right) \right), \quad (52)$$

$$v_{n+1} = \max \left(0, \min \left(v_{\text{free}}, v_{c,n}^{(\text{ACC})}, v_{s,n} \right) \right), \quad (53)$$

$\lfloor z \rfloor$ denotes the integer part of z . Through the use of formula (52), acceleration and deceleration of the ACC-vehicles are limited by some maximum acceleration a_{max} and maximum deceleration

$b_{\max}$, respectively. Owing to the formula (53), the speed of the ACC-vehicle v_{n+1} at time step $n+1$ is limited by the maximum speed in free flow v_{free} and by the safe speed $v_{s,n}$ to avoid collisions between vehicles. The maximum speed in free flow v_{free} and the safe speed $v_{s,n}$ are chosen, respectively, the same as those in the microscopic model of human driving vehicles (Appendix A). It should be noted that the model of ACC-vehicle merging from the on-ramp onto the main road is similar to that for human driving vehicles (see Appendix D).

Simulations show that the use of the safe speed in formula (53) does not influence on the dynamics of the ACC-vehicles (4) in *free flow* outside the bottleneck. However, due to vehicle merging from the on-ramp onto the main road, time headway of the vehicle to the preceding vehicle can be considerably smaller than $\tau_d^{(\text{ACC})}$. Therefore, formula (53) allows us to avoid collisions of the ACC-vehicle with the preceding vehicle in such dangerous situations. Moreover, very small values of time headway can occur in congested traffic; formula (53) prevents vehicle collisions in these cases also.

Appendix C. Discrete Version of TPACC Model

The model for human driving vehicles (Kerner and Klenov 2002, 2003, 2009) used in all simulations is discrete in time (Appendix A). Therefore, we simulate TPACC model (9) with discrete time $t_n = \tau n$, $n = 0, 1, \dots$. Respectively, TPACC model (9) should be rewritten as follows:

$$a_n^{(\text{TPACC})} = \begin{cases} K_{\Delta v}(v_{\ell,n} - v_n) & \text{at } g_n \leq G_n \\ K_1(g_n - v_n \tau_p) + K_2(v_{\ell,n} - v_n) & \text{at } g_n > G_n, \end{cases} \quad (54)$$

where $G_n = v_n \tau_G$ and $\tau_p < \tau_G$.

Safety Conditions

When $g_n < g_{\text{safe},n}$, the TPACC-vehicle should move in accordance with some safety conditions to avoid collisions between vehicles (Fig. 3).

A collision-free TPACC-vehicle motion is described as made in (Kerner 2016) for the classical model of ACC:

$$v_{c,n}^{(\text{TPACC})} = v_n + \tau \max\left(-b_{\max}, \min\left(\left\lfloor a_n^{(\text{TPACC})} \right\rfloor, a_{\max}\right)\right), \quad (55)$$

$$v_{n+1} = \max\left(0, \min\left(v_{\text{free}}, v_{c,n}^{(\text{TPACC})}, v_{s,n}\right)\right), \quad (56)$$

where the TPACC acceleration and deceleration are limited by $a_{\max}$ and $b_{\max}$, respectively; the speed v_{n+1} (56) at time step $n+1$ is limited by the maximum speed v_{free} and by the safe speed $v_{s,n}$ that have been chosen, respectively, the same as those in the model of human driving vehicles; $\lfloor z \rfloor$ denotes the integer part of z .

“Indifference Zone” in Car-Following

In accordance with Eq. (56), condition $v_{c,n}^{(\text{TPACC})} \leq v_{s,n}$ is equivalent to condition $g_n \geq g_{\text{safe},n}$. Under this condition, from the TPACC model (54)–(56), it follows that when time headway $\tau_n^{(\text{net})} = g_n/v_n$ of the TPACC-vehicle to the preceding vehicle is within the range

$$\tau_{\text{safe},n} \leq \tau_n^{(\text{net})} \leq \tau_G, \quad (57)$$

the acceleration (deceleration) of the TPACC-vehicle does not depend on time headway. In (57), $\tau_{\text{safe},n} = g_{\text{safe},n}/v_n$ is a safe time headway, and it is assumed that $v_n > 0$.

In accordance with (57), for the TPACC model (54)–(56), there is *no* fixed desired time headway to the preceding vehicle (Fig. 3). This means that in the TPACC model (54)–(56), there is “indifference zone” in the choice of time headway in car-following. This is in contrast with the classical ACC model (4) for which there is a fixed desired time headway in car-following. It should be noted that formula (57) for the indifference zone in time headway of TPACC is a discrete version of formula (13) for the indifference zone in time headway of TPACC discussed in the main text.

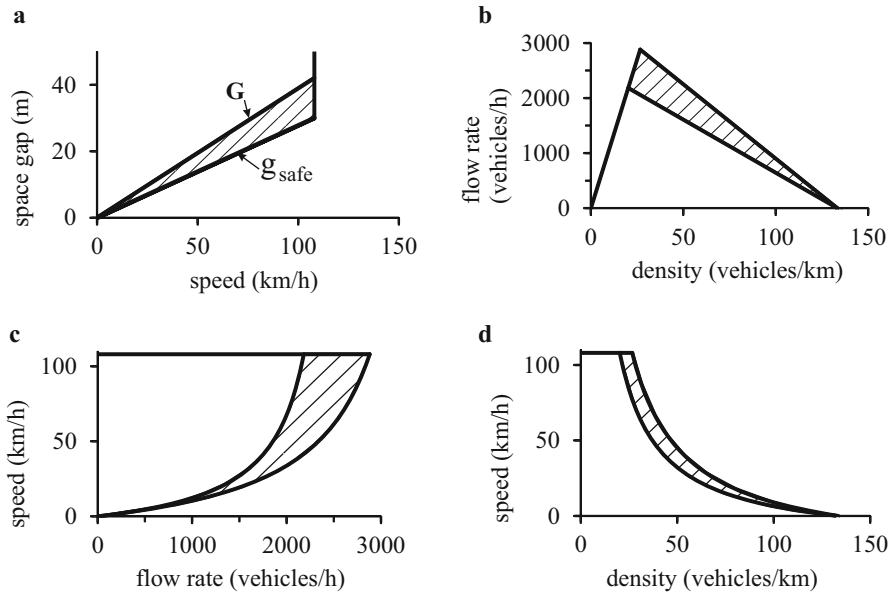
Operating Points

From formula for the safe speed $v_{s,n}$ in (56) that is given in Appendix A, we find that the safe time headway $\tau_{\text{safe},n}$ in (13) for the operating points of TPACC model (54), (55), and (56) is a constant value that is equal to $\tau_{\text{safe}} = \tau = 1$ s. In operating points of TPACC model (54), (55), and (56), $a^{(\text{TPACC})} = 0$; respectively, $v = v_{\text{free}}$ $g_{\text{safe}}(v) \leq g \leq G(v)$, and $v = v_{\text{free}}$ at $g > G(v)$, where $g_{\text{safe}}(v) = v\tau$ and $G(v) = v\tau_G$. The operating points of the TPACC model (54), (55), and (56) cover a 2D region in the space gap–speed plane (dashed 2D region in Fig. 19a). The inequalities $v \leq v_{\text{free}}$, $g \leq G(v)$, and $g \geq g_{\text{safe}}(v)$ define a 2D region in the space gap–speed plane (Fig. 19a) in which the operating points exist for the discrete version of TPACC model (54), (55), and (56).

It should be noted that the speed v and space gap g are integer in the discrete version of TPACC model (54), (55), and (56). Therefore, the operating points do not form a continuum in the space gap–speed plane as they do in the continuum version of TPACC model (9).

From Fig. 19, we can see that under conditions $g_{\text{safe}}(v) \leq g \leq G(v)$ for each given speed $v > 0$ of TPACC, there is no fixed time headway to the preceding vehicle in operating points of the TPACC model (dashed 2D regions in Fig. 19), as explained in section “ACC in Framework of Three-Phase Theory (TPACC)” (Fig. 3).

The discretization interval of TPACC acceleration (deceleration) made in TPACC model (54), (55), and (56) is chosen to be an extremely small value that is equal to $\delta a = 0.01 \text{ m/s}^2$ (see Appendix A). Therefore, the maximum value of a small round down of $a_n^{(\text{TPACC})}$ in Eqs. (55) and (56) through the application of the floor operator $\lfloor a_n^{(\text{TPACC})} \rfloor$ is less than 0.01 m/s^2 , and it is, therefore, negligible. We have tested that *no* conclusions about physical features of TPACC dynamic behavior have been changed, when the continuum in space model of human driving vehicles of Ref. (Kerner and Klenov 2003) and, respectively, the continuum in space TPACC model version (without the floor operator) are used (the reason for the use of the discrete in space model for human



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 19 Operating points of TPACC model (54)–(56) presented in the space gap–speed (a), flow–density (b), speed–flow (c), and speed–density

(d) planes. Model parameters $\tau_G = 1.4$ s, $\tau_p = 1.3$ s, $\tau_{\text{safe}} = 1$ s, $v_{\text{free}} = 30$ m/s (108 km/h), vehicle length (including the mean space gap between vehicles stopped within a wide moving jam) $d = 7.5$ m

driving vehicles of Ref. (Kerner and Klenov 2009) that leads to Eqs. (55) and (56) has been explained in (Kerner and Klenov 2009) as well as in Appendix A of the book (Kerner 2017a).

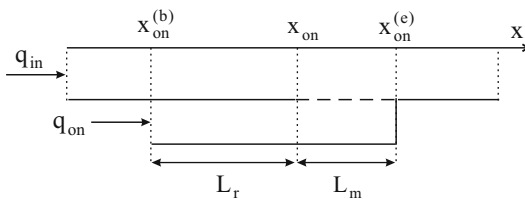
Simulations show that the use of the safe speed in formula (56) does not influence on the dynamics of the TPACC-vehicles (54) in free flow outside the bottleneck. However, formula (56) allows us to avoid collisions of the TPACC-vehicle with the preceding vehicle in dangerous situations that can occur at the bottleneck as well as in congested traffic.

Appendix D. Model of On-Ramp Bottleneck

An on-ramp bottleneck consists of two parts (Fig. 20):

- (i) The merging region of length L_m where vehicle can merge onto the main road from the on-ramp lane.
- (ii) A part of the on-ramp lane of length L_r upstream of the merging region where vehicles move in accordance with the model of Appendix A. The maximal speed of vehicles is $v_{\text{free}} = v_{\text{free on}}$.

At the beginning of the on-ramp lane ($x = x_{\text{on}}^{(b)}$), the flow rate to the on-ramp q_{on} is given through boundary conditions that are the same as those that determine the flow rate q_{in} at the beginning of the main road (see Appendix E below).



Autonomous Driving in the Framework of Three-Phase Traffic Theory, Fig. 20 Model of on-ramp bottleneck on single-lane road

Model of Vehicle Merging at Bottleneck

Vehicle Speed Adaptation Within Merging Region of Bottleneck

For the on-ramp bottleneck, when a vehicle is within the merging region of the bottleneck, the vehicle takes into account the space gaps to the preceding vehicles and their speeds both in the current and target lanes. Respectively, instead of formula (19), in (18) for the speed $v_{c,n}$, the following formula is used:

$$v_{c,n} = \begin{cases} v_n + \Delta_n^+ & \text{at } g_n^+ \leq G(v_n, \hat{v}_n^+) \\ v_n + a_n \tau & \text{at } g_n^+ > G(v_n, \hat{v}_n^+), \end{cases} \quad (58)$$

$$\Delta_n^+ = \max(-b_n \tau, \min(a_n \tau, \hat{v}_n^+ - v_n)), \quad (59)$$

$$\hat{v}_n^+ = \max(0, \min(v_{\text{free}}, v_n^+ + \Delta v_r^{(2)})), \quad (60)$$

$\Delta v_r^{(2)}$ is constant (see Table 2).

Superscripts $+$ and $-$ in variables, parameters, and functions denote the preceding vehicle and the trailing vehicle in the “target” (neighboring) lane, respectively. The target lane is the lane into which the vehicle wants to change.

The safe speed $v_{s,n}$ in (16) and (18) for the vehicle that is the closest one to the end of the merging region is chosen in the form

$$v_{s,n} = \left[v^{(\text{safe})} (x_{\text{on}}^{(e)} - x_n, 0) \right] \quad (61)$$

(see Table 2).

Safety Conditions for Vehicle Merging

Vehicle merging at the bottleneck occurs, when safety conditions (*) or safety conditions (**) are satisfied.

Autonomous Driving in the Framework of Three-Phase Traffic Theory, Table 2 Parameters of model of on-ramp bottleneck used in simulations of the main text

$$\lambda_b = 0.75, v_{\text{free on}} = 22.2 \text{ ms}^{-1}/\delta v,$$

$$\Delta v_r^{(2)} = 5 \text{ ms}^{-1}/\delta v,$$

$$L_r = 1 \text{ km}/\delta x, \Delta v_r^{(1)} = 10 \text{ ms}^{-1}/\delta v,$$

$$L_m = 0.3 \text{ km}/\delta x.$$

Safety conditions (*) are as follows:

$$\begin{aligned} g_n^+ &> \min(\hat{v}_n \tau, G(\hat{v}_n, v_n^+)), \\ g_n^- &> \min(v_n^- \tau, G(v_n^-, \hat{v}_n)), \end{aligned} \quad (62)$$

$$\hat{v}_n = \min(v_n^+, v_n + \Delta v_r^{(1)}), \quad (63)$$

$\Delta v_r^{(1)} > 0$ is constant (see Fig. 20 and Table 2).

Safety conditions (**) are as follows:

$$x_n^+ - x_n^- - d > g_{\text{target}}^{(\min)}, \quad (64)$$

where

$$g_{\text{target}}^{(\min)} = \lfloor \lambda_b v_n^+ + d \rfloor, \quad (65)$$

λ_b is constant. In addition to conditions (64), the safety condition (**) includes the condition that the vehicle should pass the midpoint

$$x_n^{(m)} = \lfloor (x_n^+ + x_n^-)/2 \rfloor \quad (66)$$

between two neighboring vehicles in the target lane, i.e., conditions

$$\begin{aligned} x_{n-1} &< x_{n-1}^{(m)} \text{ and } x_n \geq x_n^{(m)} \\ \text{or} \\ x_{n-1} &\geq x_{n-1}^{(m)} \text{ and } x_n < x_n^{(m)}. \end{aligned} \quad (67)$$

should also be satisfied.

Speed and Coordinate of Vehicle after Vehicle Merging

The vehicle speed after vehicle merging is equal to

$$v_n = \hat{v}_n. \quad (68)$$

Under conditions (*), the vehicle coordinates x_n remains the same. Under conditions (**), the vehicle coordinates x_n is equal to

$$x_n = x_n^{(m)}. \quad (69)$$

Merging of ACC-Vehicle or TPACC-Vehicle at On-Ramp Bottleneck

Here we consider rules of the merging of an ACC-vehicle at the on-ramp bottleneck presented in (Kerner 2017a) and used in simulations. The same rules have also been used in simulations of

the merging of a TPACC-vehicle from the on-ramp lane onto the main road at the bottleneck.

In the on-ramp lane, an ACC-vehicle or a TPACC-vehicle moves in accordance with the ACC model (51), (52), and (53) or in accordance with the TPACC model (54), (55), and (56), respectively. The maximal speed of the ACC-vehicle or the TPACC-vehicle in the on-ramp lane is $v_{\text{free}} = v_{\text{free on}}$. The safe speed $v_{s,n}$ in (53) for the ACC-vehicle and in (56) for the TPACC-vehicle that is the closest one to the end of the merging region is the same as that for human driving vehicles that is given by formula (61).

An ACC-vehicle or a TPACC-vehicle merges from the on-ramp lane onto the main road, when some safety conditions (*) or safety conditions (**) are satisfied for the ACC-vehicle or the TPACC-vehicle. Safety conditions (*) for ACC-vehicles and TPACC-vehicles are as follows:

$$g_n^+ > \hat{v}_n \tau, \quad g_n^- > v_n^- \tau, \quad (70)$$

where $\hat{v}_n$ is given by formula (63). Safety conditions (**) are given by formulas (64), (65), (66), and (67), i.e., they are the same as those for human driving vehicles. Respectively, as for human driving vehicles, the ACC-vehicle speed and its coordinate or the TPACC-vehicle speed and its coordinate after the ACC-vehicle or the TPACC-vehicle has merged from the on-ramp onto the main road are determined by formulas (68) and (69).

A question can arise from the choice of the model time step $\tau = 1$ s in Eqs. (54), (55), and (56) that have been used for numerical simulations of TPACC model (9): In TPACC model (54), (55), and (56), the time step $\tau = 1$ s determines the safe space gap $g_{\text{safe}} = v\tau$ under hypothetical steady-state conditions in which all vehicles move at time-independent speed v . Contrarily to the TPACC model (54), (55), and (56), typical ACC controllers in vehicles that on the market have update time intervals τ of 100 ms or less. Indeed, there may be some very dangerous traffic situations in real traffic in which the safe time headway for an ACC-vehicle is quickly reached, and, therefore, the ACC-vehicle must decelerate strongly already after a time interval that is a much shorter than 1 s to avoid the collision with the preceding vehicle. Therefore, to avoid

collisions, *real* ACC controllers must have update time intervals τ of 100 ms or less. However, at model time step $\tau = 1$ s through the choice in the mathematical formulation of the safe speed in TPACC model (54), (55), and (56) and in the model of human driving vehicles, *collision-less* traffic flow is guaranteed in *any* dangerous traffic situation that can occur in simulations of traffic flow.

In other words, to disclose the physics of TPACC it is sufficient the choice of the update time $\tau = 1$ s in *simulations* of TPACC behavior. To explain this, we should note that Eqs. (55) and (56) effect on TPACC dynamics *only* under condition $g_n < g_{\text{safe},n}$, i.e., when the space gap becomes smaller than the safe one. This is because the physics of TPACC disclosed in this entry is *solely* determined by Eq. (54): Under condition $g_n \geq g_{\text{safe},n}$, Eqs. (55) and (56) do not change TPACC acceleration (deceleration) calculated through Eq. (54).

Appendix E. Boundary Conditions for Mixed Traffic Flow

Open boundary conditions are applied. At the beginning of the road, new vehicles are generated one after another in each of the lanes of the road at time instants

$$t^{(m)} = \tau \lceil m\tau_{\text{in}}/\tau \rceil, m = 1, 2, \dots \quad (71)$$

In (71), $\tau_{\text{in}} = 1/q_{\text{in}}$, q_{in} is the flow rate in the incoming boundary flow per lane, and $\lceil z \rceil$ denotes the nearest integer greater than or equal to z . Human driving and autonomous driving vehicles are randomly generated at the beginning of the road with the rates related to chosen values of the flow rate q_{in} and the percentage γ of the autonomous driving vehicles in mixed traffic flow: (i) At time instant $t^{(m)}$ (71), a new vehicle is human driving vehicle, when condition

$$r_2 \geq \gamma/100 \quad (72)$$

is satisfied where $r_2 = \text{rand}(0, 1)$. (ii) At time instant $t^{(m)}$ (71), a new vehicle is autonomous driving vehicle, when the opposite condition

$$r_2 < \gamma/100 \quad (73)$$

is satisfied. The same procedure of random generation of human driving and autonomous driving vehicles is applied at the beginning of the on-ramp lane. In this case, however, in (71) the flow rate q_{in} should be replaced by the on-ramp inflow rate q_{on} .

A new vehicle appears on the road only if the distance from the beginning of the road ($x = x_b$) to the position $x = x_{\ell,n}$ of the farthest upstream vehicle on the road is not smaller than the safe distance $v_{\ell,n}\tau + d$:

$$x_{\ell,n} - x_b \geq v_{\ell,n}\tau + d, \quad (74)$$

where $n = t^{(m)}/\tau$. Otherwise, condition (74) is checked at time $(n+1)\tau$ that is the next one to time $t^{(m)}$ (71) and so on, until the condition (74) is satisfied. Then the next vehicle appears on the road. After this occurs, the number m in (71) is increased by 1.

The speed v_n and coordinate x_n of the new vehicle are

$$\begin{aligned} v_n &= v_{\ell,n}, \\ x_n &= \max(x_b, x_{\ell,n} - \lfloor v_n\tau_{\text{in}} \rfloor). \end{aligned} \quad (75)$$

The flow rate q_{in} is chosen to have the value $v_{\text{free}}\tau_{\text{in}}$ integer. In the initial state ($n = 0$), all vehicles have the free flow speed $v_n = v_{\text{free}}$, and they are positioned at space intervals $x_{\ell,n} - x_n = v_{\text{free}}\tau_{\text{in}}$.

After a vehicle has reached the end of the road, it is removed. Before this occurs, the farthest downstream vehicle maintains its speed and lane. For the vehicle following the farthest downstream one, the “anticipation” speed $v_{\ell}^{(a)}$ in (40) is equal to the speed of the farthest downstream vehicle.

Bibliography

- Automated Highway Systems (2007) http://www.seminarsonly.com/Civil_Engineering/automated-highway-systems.php
- Automated Highway Systems (2012) <https://seminarprojects.blogspot.de/2012/01/detailed-report-on-automated-highway.html>

- Automatisches Fahren (2012) <http://www.tuvpt.de/index.php?id=foerderung000>
- Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793–800
- Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. *Math Mod Meth App Sc* 12:1801–1843
- Bengler K, Dietmayer K, Farber B, Maurer M, Stiller C, Winner H (2014) Three decades of driver assistance systems: review and future perspectives. *IEEE Intell Transp Sys Mag* 6:6–22
- Borsche R, Kimathi M, Klar A (2012) A class of multi-phase traffic theories for microscopic, kinetic and continuum traffic models. *Comput Math Appl* 64:2939–2953
- Bose A, Ioannou P (2003) Mixed manual/semi-automated traffic: a macroscopic analysis. *Transp Res C* 11:439–462
- Brockfeld E, Kühne RD, Skabardonis A, Wagner P (2003) Toward benchmarking of microscopic traffic flow models. *Trans Res Rec* 1852:124–129
- Chen D, Ahn S, Chitturi M, Noyce DA (2017) Towards vehicle automation: roadway capacity formulation for traffic mixed with regular and automated vehicles. *Transp Res B* 100:196–221
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199–329
- Daganzo CF (1997) Fundamentals of transportation and traffic operations. Elsevier Science Inc., New York
- Davis LC (2004a) Multilane simulations of traffic phases. *Phys Rev E* 69:016108
- Davis LC (2004b) Effect of adaptive cruise control systems on traffic flow. *Phys Rev E* 69:066110
- Davis LC (2006a) Controlling traffic flow near the transition to the synchronous flow phase. *Phys A* 368:541–550
- Davis LC (2006b) Effect of cooperative merging on the synchronous flow phase of traffic. *Phys A* 361:606–618
- Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. *Phys A* 379:274–290
- Davis LC (2008) Driver choice compared to controlled diversion for a freeway double on-ramp in the framework of three-phase traffic theory. *Phys A* 387:6395–6410
- Davis LC (2014) Nonlinear dynamics of autonomous vehicles with limits on acceleration. *Phys A* 405:128–139
- Davis LC (2016) Improving traffic flow at a 2-to-1 lane reduction with wirelessly connected, adaptive cruise control vehicles. *Phys A* 451:320–332
- Delis AI, Nikolos IK, Papageorgiou M (2015) Macroscopic traffic flow modeling with adaptive cruise control: development and numerical solution. *Comput Math Appl* 70:1921–1947
- Dharba S, Rajagopal KR (1999) Intelligent cruise control systems and traffic flow stability. *Transp Res C* 7:329–352
- Elefteriadou L (2014) An introduction to traffic flow theory. Springer optimization and its applications, vol 84. Springer, Berlin
- Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Transp Res Rec* 1484:80–89
- European Roadmap Smart Systems for Automated Driving (2015) <https://www.smart-systems-integration.org/public/documents/>
- Ferrara A, Saccone S, Siri S (2018) Freeway traffic modeling and control. Springer, Berlin. <https://doi.org/10.1007/978-3-319-75961-6>
- Gao K, Jiang R, Hu S-X, Wang B-H, Wu Q-S (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. *Phys Rev E* 76:026105
- Gao K, Jiang R, Wang B-H, Wu Q-S (2009) Discontinuous transition from free flow to synchronized flow induced by short-range interaction between vehicles in a three-phase traffic flow model. *Phys A* 388:3233–3243
- Gartner NH, Messer CJ, Rathi A (eds) (1997) Special report 165: revised monograph on traffic flow theory. Transportation Research Board, Washington, DC
- Gartner NH, Messer CJ, Rathi A (eds) (2001) Traffic flow theory: a state-of-the-art report. Transportation Research Board, Washington, DC
- Gasnikov AV, Klenov SL, Nurminski EA, Kholodov YA, Shamray NB (2013) Introduction to mathematical simulations of traffic flow. MCNMO, Moscow (in Russian)
- Gazis DC (2002) Traffic theory. Springer, Berlin
- Gipps PG (1981) Behavioral car-following model for computer simulation. *Trans Res B* 15:105–111
- Haight FA (1963) Mathematical theories of traffic flow. Academic, New York
- Han Y, Ahn S (2018) Stochastic modeling of breakdown at freeway merge bottleneck and traffic control method using connected automated vehicle. *Transp Res B* 107:146–166
- Hausken K, Rehborn H (2015) Game-theoretic context and interpretation of Kerner's three-phase traffic theory. In: Hausken K, Zhuang J (eds) Game theoretic analysis of congestion, safety and security. Springer series in reliability engineering. Springer, Berlin, pp 113–141
- He S, Guan W, Song L (2010) Explaining traffic patterns at on-ramp vicinity by a driver perception model in the framework of three-phase traffic theory. *Phys A* 389:825–836
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Helbing D, Hennecke A, Treiber M (1999) Phase diagram of traffic states in the presence of inhomogeneities. *Phys Rev Lett* 82:4360–4363
- Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) Traffic and granular flow' 99. Springer, Heidelberg
- Helbing D, Treiber M, Kesting A, Schönhof M (2009) Theoretical vs. empirical classification and prediction of congested traffic states. *Eur Phys J B* 69:583–598
- Highway Capacity Manual (2000) National research council. Transportation Research Board, Washington, DC
- Highway Capacity Manual (2010) National research council. Transportation Research Board, Washington, DC
- Hoogendoorn S, van Lint H, Knoop VL (2008) Macroscopic modeling framework unifying kinematic wave modeling and three-phase traffic theory. *Trans Res Rec* 2088:102–108

- Ioannou PA (ed) (1997) Automated highway systems. Plenum Press, New York
- Ioannou P, Chien CC (1993) Autonomous Intelligent Cruise Control. *IEEE Trans Veh Technol* 42:657–672
- Ioannou PA, Kosmatopoulos EB (2000) Adaptive control. In: Webster JG (ed) Wiley encyclopedia of electrical and electronics engineering. Wiley, New York. <https://doi.org/10.1002/047134608X.W1002>
- Ioannou PA, Sun J (1996) Robust adaptive control. Prentice Hall, Inc., Upper Saddle River
- Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. *J Phys A Math Gen* 37:8197–8213
- Jiang R, Wu QS (2005) Toward an improvement over Kerner-Klenov-Wolf three-phase cellular automaton model. *Phys Rev E* 72:067103
- Jiang R, Wu QS (2007a) Dangerous situations in a synchronized flow model. *Phys A* 377:633–640
- Jiang R, Wu QS (2007b) Dangerous situations in a synchronized flow model. *Phys A* 377:633–640
- Jiang R, Hu M-B, Wang R, Wu Q-S (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner-Klenov speed adaptation model. *Phys Lett A* 365:6–9
- Jiang R, Hu MB, Zhang HM, Gao ZY, Jia B, Wu QS, Yang M (2014) Traffic experiment reveals the nature of car-following. *PLoS One* 9:e94351
- Jiang R, Hu M-B, Zhang HM, Gao ZY, Jia B, Wu QS (2015) On some experimental features of car-following behavior and how to model them. *Transp Res B* 80:338–354
- Jiang R, Jin C-J, Zhang HM, Huang Y-X, Tian J-F, Wang W, Hu M-B, Wang H, Jia B (2017) Experimental and empirical investigations of traffic flow instability. *Transp Res Proc* 23:157–173
- Jin C-J, Wang W (2011) The influence of nonmonotonic synchronized flow branch in a cellular automaton traffic flow model. *Phys A* 390:4184–4191
- Jin C-J, Wang W, Jiang R, Gao K (2010) On the first-order phase transition in a cellular automaton traffic flow model without a slow-to-start effect. *J Stat Mech* 2010:P03018
- Jin C-J, Wang W, Jiang R, Zhang HM, Wang H, Hu M-B (2015) Understanding the structure of hyper-congested traffic from empirical and experimental evidences. *Transp Res C* 60:324–338
- Kerner BS (1998a) Traffic flow: experiment and theory. In: Schreckenberg M, Wolf DE (eds) Traffic and granular flow'97. Springer, Singapore, pp 239–267
- Kerner BS (1998b) Theory of congested traffic flow. In: Rysgaard R (ed) Proceedings of the 3rd symposium on highway capacity and level of service, vol 2. Road Directorate, Ministry of Transport, Denmark, pp 621–642
- Kerner BS (1998c) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
- Kerner BS (1999a) Congested traffic flow: observations and theory. *Trans Res Rec* 1678:160–167
- Kerner BS (1999b) Theory of congested traffic flow: self-organization without bottlenecks. In: Ceder A (ed) Transportation and traffic theory. Elsevier Science, Amsterdam, pp 147–171
- Kerner BS (1999c) The physics of traffic. *Phys World* 12:25–30
- Kerner BS (2000) Experimental features of the emergence of moving jams in free traffic flow. *J Physics A: Math Gen* 33:L221–L228
- Kerner BS (2001) Complexity of synchronized flow and related problems for basic assumptions of traffic flow theories. *Netw Spat Econ* 1:35–76
- Kerner BS (2002a) Synchronized flow as a new traffic phase and related problems for traffic flow modelling. *Math Comput Model* 35:481–508
- Kerner BS (2002b) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
- Kerner BS (2004) The physics of traffic. Springer, Berlin
- Kerner BS (2007a) Method for actuating a traffic-adaptive assistance system which is located in a vehicle, USA patent US 20070150167A1. <https://google.com/patents/US20070150167A1>; USA patent US 7451039B2 (2008)
- Kerner BS (2007b) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assistenzsystem, German patent publication DE 102007008253A1. <https://register.dpma.de/DPMAREgister/pat/PatSchrifteneinsicht?docId=DE102007008253A1>
- Kerner BS (2007c) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assistenzsystem, German patent publication DE 102007008257A1. <https://register.dpma.de/DPMAREgister/pat/PatSchrifteneinsicht?docId=DE102007008257A1>
- Kerner BS (2008) Betriebsverfahren für ein fahrzeugseitiges verkehrsadaptives Assisten-system, German patent publication DE 102007008254A1
- Kerner BS (2009) Introduction to modern traffic flow theory and control. Springer, Berlin
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Phys A* 392:5261–5282
- Kerner BS (2015a) Microscopic theory of traffic-flow instability governing traffic breakdown at highway bottlenecks: growing wave of increase in speed in synchronized flow. *Phys Rev E* 92:062827
- Kerner BS (2015b) Failure of classical traffic flow theories: a critical review. *Elektrotechnik und Informationstechnik* 132:417–433
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Phys A* 450:700–747
- Kerner BS (2017a) Breakdown in traffic networks: fundamentals of transportation science. Springer, Berlin
- Kerner BS (2017b) Traffic networks, breakdown in. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer Science+Business Media LLC, Springer, Berlin. <https://doi.org/10.1007/978-3-642-27737-5701-1>

- Kerner BS (2017c) Physics of autonomous driving based on three-phase traffic theory. arXiv:1710.10852v3. <http://arxiv.org/abs/1710.10852>
- Kerner BS (2017d) Traffic breakdown, modeling approaches to. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer Science+Business Media LLC, Springer, Berlin. <https://doi.org/10.1007/978-3-642-27737-5559-2>
- Kerner BS (2018a) Traffic congestion, spatiotemporal features of. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer Science+Business Media LLC, Springer, Berlin
- Kerner BS (2018b) Physics of automated driving in framework of three-phase traffic theory. Phys Rev E 97:042303
- Kerner BS (2018c) Autonomous driving in framework of three-phase traffic theory. Procedia Comput Sci 130:785–790. <https://doi.org/10.1016/j.procs.2018.04.136>
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. J Phys A Math Gen 35: L31–L43
- Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. Phys Rev E 68:036130
- Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. J Phys A Math Gen 39:1775–1809
- Kerner BS, Klenov SL (2009) Phase transitions in traffic flow on multilane roads. Phys Rev E 80:056101
- Kerner BS, Klenov SL (2018) Traffic breakdown, mathematical probabilistic approaches to. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer Science+Business Media LLC, Springer, Berlin
- Kerner BS, Rehborn H (1996) Experimental properties of complexity in traffic flow. Phys Rev E 53:R4275–R4278
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. Phys Rev Lett 79:4030–4033
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. J Phys A Math Gen 35:9971–10013
- Kerner BS, Klenov SL, Hiller A (2006a) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. J Phys A Math Gen 39:2001–2020
- Kerner BS, Klenov SL, Hiller A, Rehborn H (2006b) Microscopic features of moving traffic jams. Phys Rev E 73:046107
- Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. Non Dyn 49:525–553
- Kerner BS, Klenov SL, Schreckenberg M (2011) Simple cellular automaton model for traffic breakdown, highway capacity, and synchronized flow. Phys Rev E 84:046110
- Kerner BS, Klenov SL, Hermanns G, Schreckenberg M (2013a) Effect of driver overacceleration on traffic breakdown in three-phase cellular automaton traffic flow models. Phys A 392:4083–4105
- Kerner BS, Rehborn H, Schäfer R-P, Klenov SL, Palmer J, Lorkowski S, Witte N (2013b) Traffic dynamics in empirical probe vehicle data studied with three-phase theory: spatiotemporal reconstruction of traffic phases and generation of jam warning messages. Phys A 392:221–251
- Kerner BS, Klenov SL, Schreckenberg M (2014) Probabilistic physical characteristics of phase transitions at highway bottlenecks: incommensurability of three-phase and two-phase traffic-flow theories. Phys Rev E 89:052807
- Kerner BS, Koller M, Klenov SL, Rehborn H, Leibel M (2015) The physics of empirical nuclei for spontaneous traffic breakdown in free flow at highway bottlenecks. Phys A 438:365–397
- Kesting A, Treiber M, Schönhof M, Helbing D (2007) Extending adaptive cruise control to adaptive driving strategies. Transp Res Rec 2000:16–24
- Kesting A, Treiber M, Schönhof M, Helbing D (2008) Adaptive cruise control design for active congestion avoidance. Transp Res C 16:668–683
- Kesting A, Treiber M, Helbing D (2010) Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. Phil Trans Royal Society Series A 368:4585–4605
- Klenov SL (2010) Kerner's three-phase traffic theory – a new theoretical basis for development of intelligent transportation systems. In: Kozlov VV (Ed) Proceedings of Moscow institute of physics and technology (State University), vol 2, pp 75–90 (in Russian)
- Knor F, Schreckenberg M (2013) The comfortable driving model revisited: traffic phases and phase transitions. J Stat Mech 2013:P07002
- Kokubo S, Tanimoto J, Hagishima A (2011) A new cellular automata model including a decelerating damping effect to reproduce Kerner's three-phase theory. Phys A 390:561–568
- Krauß S (1998) Microscopic modeling of traffic flow: investigation of collision free vehicle dynamics. Ph.D. thesis, University of Cologne, Germany. <http://e-archive.informatik.uni-koeln.de/319/>
- Krauß S, Wagner P, Gawron C (1997) Metastable states in a microscopic model of traffic flow. Phys Rev E 55:5597–5602
- Kuhn TS (2012) The structure of scientific revolutions, 4th edn. The University of Chicago Press, Chicago/London
- Kukuchi S, Uno N, Tanaka M (2003) Impacts of shorter perception-reaction time of adapted cruise controlled vehicles on traffic flow and safety. Transp Eng 129:146–154
- Lee HK, Kim B-J (2011) Dissolution of traffic jam via additional local interactions. Phys A 390:4555–4561
- Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical restriction versus human overreaction triggering congested traffic states. Phys Rev Lett 92:238702
- Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin

- Levine W, Athans M (1966) On the optimal error regulation of a string of moving vehicles. *IEEE Trans Automat Contr* 11:355–361
- Li PY, Shrivastava A (2002) Traffic flow stability induced by constant time headway policy for adaptive cruise control vehicles. *Transp Res C* 10:275–301
- Li XG, Gao ZY, Li KP, Zhao XM (2007) Relationship between microscopic dynamics in traffic flow and complexity in networks. *Phys Rev E* 76:016110
- Liang C-Y, Peng H (1999) Optimal adaptive cruise control with guaranteed string stability. *Veh Syst Dyn* 32:313–330
- Liang C-Y, Peng H (2000) String stability analysis of adaptive cruise controlled vehicles. *JSME Int J Ser C* 43:671–677
- Lin T-W, Hwang S-L, Green P (2009) Effects of time-gap settings of adaptive cruise control (ACC) on driving performance and subjective acceptance in a bus driving simulator. *Saf Sci* 47:620–625
- Mahnke R, Kaupuzs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
- Mahnke R, Kaupuzs J, Lubashevsky I (2009) Physics of stochastic processes: how randomness acts in time. Wiley-VCH, Weinheim
- Mamouei M, Kaparias I, Halikias G (2018) A framework for user- and system-oriented optimisation of fuel efficiency and traffic flow in Adaptive Cruise Control. *Transp Res C* 92:27–41
- Marsden G, McDonald M, Brackstone M (2001) Towards an understanding of adaptive cruise control. *Transp Res C* 9:33–51
- Martinez J-J, Canudas-do-Wit C (2007) A safe longitudinal control for adaptive cruise control and stop-and-go scenarios. *IEEE Trans Control Syst Technol* 15:246–258
- Maurer M, Gerdes JC, Lenz B, Winner H (eds) (2015) *Autonomes Fahren*. Springer, Berlin
- May AD (1990) *Traffic flow fundamentals*. Prentice-Hall, Inc., New Jersey
- Meyer G, Beiker S (2014) *Road vehicle automation*. Springer, Berlin
- Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
- Nagel K, Wagner P, Woesler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–716
- Neto JPL, Lyra ML, da Silva CR (2011) Phase coexistence induced by a defensive reaction in a cellular automaton traffic flow model. *Phys A* 390:3558–3565
- Newell GF (1982) *Applications of queuing theory*. Chapman Hall, London
- Ngoduy D (2012) Application of gas-kinetic theory to modelling mixed traffic of manual and ACC vehicles. *Transpmetrics* 8:43–60
- Ngoduy D (2013) Instability of cooperative adaptive cruise control traffic flow: a macroscopic approach. *Commun Nonlinear Sci Numer Simul* 18:2838–2851
- Ntousakis IA, Nokolos IK, Papageorgiou M (2015) On microscopic modelling of adaptive cruise control systems. *Transp Res Procedia* 9:111–127
- Papageorgiou M (1983) *Application of automatic control concepts in traffic flow modeling and control*. Springer, Berlin
- Perraki G, Roncoli C, Papamichail I, Papageorgiou M (2018) Evaluation of a model predictive control framework for motorway traffic involving conventional and automated vehicles. *Trans Res C* 92:456–471
- Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Trans Res Rec* 1634:64–69
- Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical restriction versus human overreaction: accident avoidance and two-lane simulations. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Traffic and granular flow'05*. Proceedings of the international workshop on traffic and granular flow. Springer, Berlin, pp 503–508
- Qian Y-S, Feng X, Jun-Wei Zeng J-W (2017) A cellular automata traffic flow model for three-phase theory. *Phys A* 479:509–526
- Rajamani R (2012) *Vehicle dynamics and control, mechanical engineering series*. Springer US, Boston
- Rehborn H, Klenov SL (2009) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9500–9536
- Rehborn H, Koller M (2014) A study of the influence of severe environmental conditions on common traffic congestion features. *J Adv Transp* 48:1107–1120
- Rehborn H, Palmer J (2008) ASDA/FOTO based on Kerner's three-phase traffic theory in north Rhine-Westphalia and its integration into vehicles. In: *Intelligent vehicles symposium, IEEE, Eindhoven, Netherlands*, pp 186–191. ISSN: 1931-0587. <https://doi.org/10.1109/IVS.2008.4621192>
- Rehborn H, Klenov SL, Palmer J (2011a) An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. *Phys A* 390:4466–4485
- Rehborn H, Klenov SL, Palmer J (2011b) Common traffic congestion features studied in USA, UK, and Germany based on Kerner's three-phase traffic theory. In: *IEEE intelligent vehicles symposium (IV), IEEE, Baden-Baden, Germany*, pp 19–24. <https://doi.org/10.1109/IVS.2011.5940394>
- Rehborn H, Klenov SL, Koller M (2017) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC, Springer, Berlin
- Rempe F, Franek P, Fastenrath U, Bogenberger K (2016) Online freeway traffic estimation with real floating car data. In: *Proceedings of 2016 IEEE 19th international conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro*, pp 1838–1843
- Roncoli C, Papageorgiou M, Papamichail I (2015) Traffic flow optimisation in presence of vehicle automation and communication systems – part I: a first-order multi-lane model for motorway traffic. *Transp Res C* 57:241–259

- Saifuzzaman M, Zheng Z (2014) Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transp Res C* 48:379–403
- Schadschneider A, Chowdhury D, Nishinari K (2011) Stochastic transport in complex systems. Elsevier Science Inc., New York
- Sharon G, Levin MW, Hanna JP, Rambha T, Boyles SD, Stone P (2017) Network- wide adaptive tolling for connected and automated vehicles. *Transp Res C* 84:142–157
- Shladover SE (1995) Review of the state of development of advanced vehicle control systems (AVCS). *Veh Syst Dyn* 24:551–595
- Shladover SE, Su D, Lu X-T (2012) Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transp Res Rec* 2324:63–70
- Siebel F, Mauser W (2006) Synchronized flow and wide moving jams from balanced vehicular traffic. *Phys Rev E* 73:066108
- Suzuki H (2003) Effect of adaptive cruise control (ACC) on traffic throughput: numerical example on actual freeway corridor. *JSAE Rev* 24:403–410
- Swaroop D, Hedrick JK (1996) String stability for a class of nonlinear systems. *IEEE Trans Automat Contr* 41:349–357
- Swaroop D, Hedrick JK, Choi SB (2001) Direct adaptive longitudinal control of vehicle platoons. *IEEE Trans Veh Technol* 50:150–161
- Takayasu M, Takayasu H (1993) Phase transition and 1/f type noise in one dimensional asymmetric particle dynamics. *Fractals* 1:860–866
- Talebpoor A, Mahmassani HS (2016) Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transp Res C* 71:143–163
- Tian J-F, Jia B, Li X-G, Jiang R, Zhao X-M, Gao Z-Y (2009) Synchronized traffic flow simulating with cellular automata model. *Phys A* 388:4827–4837
- Tian J-F, Yuan Z-Z, Jia B, Treiber M, Jia B, Zhang W-Y (2012) Cellular automaton model within the fundamental-diagram approach reproducing some findings of the three-phase theory. *Phys A* 391:3129–3139
- Tian J-F, Jiang R, Jia B, Gao Z-Y, Ma SF (2016a) Empirical analysis and simulation of the concave growth pattern of traffic oscillations. *Transp Res B* 93:338–354
- Tian J-F, Jiang R, Li G, Treiber M, Jia B, Zhu CQ (2016b) Improved 2D intelligent driver model in the framework of three-phase traffic theory simulating synchronized flow and concave growth pattern of traffic oscillations. *Transp Res F* 41:55–65
- Tian J-F, Li G, Treiber M, Jiang R, Jia N, Ma SF (2016c) Cellular automaton model simulating spatiotemporal patterns, phase transitions and concave growth pattern of oscillations in traffic flow. *Transp Res B* 93:560–575
- Tian J-F, Zhu C-Q, Jiang R (2018) Cellular automaton models in the framework of three-phase traffic theory. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC, Springer, Berlin
- Treiber M, Helbing D (2001) Microsimulations of freeway traffic including control measures. *Automatisierungstechnik* 49:478–484
- Treiber M, Kesting A (2013) *Traffic flow dynamics*. Springer, Berlin
- van Arem B, van Driel CJG, Visser R (2006) The impact of cooperative adaptive cruise control on traffic flow characteristics. *IEEE Trans Intell Transp Syst* 7:429–436
- Van Brummelen J, O'Brien M, Gruyer D, Najjaran H (2018) Autonomous vehicle perception: the technology of today and tomorrow. *Transp Res C* 89:384–406
- VanderWerf J, Shladover SE, Kourjanskaia N, Miller M, Krishnan H (2001) Modeling effects of driver control assistance systems on traffic. *Transp Res Rec* 1748:167–174
- VanderWerf J, Shladover SE, Miller MA, Kourjanskaia N (2002) Effects of adaptive cruise control systems on highway traffic flow capacity. *Transp Res Rec* 1800:78–84
- Varaiya P (1993) Smart cars on smart roads: problems of control. *IEEE Trans Autom Control* 38:195–207
- Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. *Phys A* 378:475–484
- Wang R, Li Y, Work DB (2017) Comparing traffic state estimators for mixed human and automated traffic flows. *Transp Res C* 78:95–110
- Whitham GB (1974) *Linear and nonlinear waves*. Wiley, New York
- Wiedemann R (1974) *Simulation des Verkehrsflusses*. University of Karlsruhe, Karlsruhe
- Wu JJ, Sun HJ, Gao ZY (2008) Long-range correlations of density fluctuations in the Kerner-Klenov-Wolf cellular automata three-phase traffic flow model. *Phys Rev E* 78:036103
- Xiang Z-T, Li Y-J, Chen Y-F, Xiong L (2013) Simulating synchronized traffic flow and wide moving jam based on the brake light rule. *Phys A* 392:5399–5413
- Yang H, Lu J, Hu X-J, Jiang J (2013) A cellular automaton model based on empirical observations of a driver's oscillation behavior reproducing the findings from Kerner's three-phase traffic theory. *Phys A* 392:4009–4018
- Yang H, Zhai X, Zheng C (2018) Effects of variable speed limits on traffic operation characteristics and environmental impacts under car-following scenarios: simulations in the framework of Kerner's three-phase traffic theory. *Phys A*. <https://doi.org/10.1016/j.physa.2018.05.032>
- Zhang P, Wu C-X, Wong SC (2012) A semi-discrete model and its approach to a solution for a wide moving jam in traffic flow. *Phys A* 391:456–463
- Zhou J, Peng H (2005) Range policy of adaptive cruise control vehicles for improved flow stability and string stability. *IEEE Trans Intell Transp Syst* 6:229–237
- Zhou M, Qu X, Jin S (2017) On the impact of cooperative autonomous vehicles in improving freeway merging: a modified intelligent driver model-based approach. *IEEE Trans Intell Transp Syst* 18:1422–1428

Spatiotemporal Features of Traffic Congestion

Boris S. Kerner

Physics of Transport and Traffic, University
Duisburg-Essen, Duisburg, Germany

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Empirical Phase Transitions in Traffic Flow

Empirical Characteristic Parameters of Wide

Moving Jam Propagation

Microscopic Features of Traffic Phases in

Congested Traffic

Moving Jam Emergence in Framework of Three-

Phase Traffic Theory

Congested Patterns at Road Bottlenecks

Congested Patterns at Heavy Bottlenecks

“Jam-Absorption” Effect in Highway and City
Traffic

Conclusions: Fundamental Empirical Features of

Spatiotemporal Congested Traffic Patterns

Future Directions

Bibliography

Glossary

Free Flow Free flow is usually observed when the vehicle density in traffic is small enough. The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function.

Congested Traffic Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is possible in free flow. In empirical observations, congested traffic occurs mostly due to traffic breakdown leading to the onset of

congestion at a road bottleneck. During traffic breakdown, the average speed decreases sharply in an initial free flow to a lower speed at the bottleneck. In highway traffic, traffic breakdown is associated with a local phase transition from free flow to synchronized flow ($F \rightarrow S$ transition). The bottleneck can be a result of roadworks, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, bad weather conditions, accidents, etc.

Effectual Bottleneck An effectual bottleneck is a bottleneck at which traffic breakdown is observed.

Three-Phase Traffic Theory The three-phase traffic theory (three-phase theory for short) is the framework for understanding of *empirical states* of traffic flow in three phases: (i) free flow, (ii) synchronized flow, and (iii) wide moving jam. The synchronized flow and wide moving jam traffic phases belong to congested traffic.

Three Traffic Phases In three-phase traffic theory, there are the three traffic phases: (i) free flow (F), (ii) synchronized flow (S), and (iii) wide moving jam (J). The wide moving jam and synchronized flow traffic phases are associated with congested traffic. The wide moving jam and synchronized flow phases of congested traffic are defined through the empirical spatiotemporal macroscopic empirical (objective) criteria [J] and [S].

S $\rightarrow$ J Instability A growth of speed disturbances of a local speed reduction in synchronized flow is called as an S $\rightarrow$ J instability. During the development of the S $\rightarrow$ J instability, growing narrow moving jams are formed. The narrow moving jam belongs to the synchronized flow phase. The emergence of narrow moving jams in synchronized flow due to the S $\rightarrow$ J instability is also called the pinch effect in synchronized flow. The S $\rightarrow$ J instability exhibits the nucleation nature. This means that small enough local disturbances in synchronized flow do

not initiate the $S \rightarrow J$ instability. To initiate the $S \rightarrow J$ instability, a local speed reduction of a large enough amplitude should appear in synchronized flow. Such a local disturbance can be considered a nucleus for the $S \rightarrow J$ instability. There can be various sources for a nucleus required for the $S \rightarrow J$ instability: unexpected braking of a vehicle in synchronized flow, lane changing on the main road leading to time headway reduction (and as a result, to local speed reduction), fluctuations in flow rates, vehicle merging from other roads, etc. During upstream propagation of growing narrow moving jams resulting from the $S \rightarrow J$ instability, some of them can transform into wide moving jams ($S \rightarrow J$ transition). Because the growth of a narrow moving jam in synchronized flow should not necessarily lead to the formation of a wide moving jam, the $S \rightarrow J$ instability should be strongly distinguished from an $S \rightarrow J$ transition.

$S \rightarrow J$ Transition Wide moving jams can emerge spontaneously in synchronized flow. The emergence of a wide moving jam in synchronized flow is called an $S \rightarrow J$ transition. The $S \rightarrow J$ transition results from the growth of a narrow moving jam (called $S \rightarrow J$ instability) that occurs in synchronized flow.

Moving Traffic Jam A moving traffic jam (moving jam for short) is a localized congested traffic pattern that moves upstream in traffic flow. Within the moving jam, the average vehicle speed is very low (sometimes as low as zero), and the density is very high. The moving jam is spatially restricted by the downstream jam front and upstream jam front. Within the downstream jam front, vehicles accelerate from low speed states within the jam to higher speeds in traffic flow downstream of the moving jam. Within the upstream jam front, vehicles must slow down to the speed within the jam.

Narrow Moving Jam A narrow moving jam is a moving jam, which consists of the jam fronts only. Narrow moving jams are associated with the synchronized flow phase. This is because there is no traffic flow interruption interval within a narrow moving jam; therefore, the

narrow moving jam does not exhibit the characteristic jam feature [J]. The narrow moving jam can be considered a state of the synchronized flow traffic phase. During the upstream propagation of a narrow moving jam, it can be surrounded upstream and downstream either by free flow or by other states of synchronized flow. A growing narrow moving jam can transform over time into a wide moving jam, i.e., the growing narrow moving jam is a nucleus for wide moving jam emergence.

Nucleus Required for Wide Moving Jam Emergence A nucleus required for wide moving jam emergence in metastable synchronized flow is a local speed (density) disturbance in the synchronized flow within which the speed is equal to or lower than the critical speed (density is equal to or larger than the critical density). A growing nucleus for wide moving jam emergence that propagates upstream in metastable synchronized flow is also called a growing narrow moving jam.

Wide Moving Jam Traffic Phase In three-phase traffic theory, the following definition (criterion) [J] of the wide moving jam traffic phase (wide moving jam for short) in congested traffic is made. The definition [J]: a wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. This is the characteristic feature [J] of the wide moving jam phase.

Spontaneous Emergence of Wide Moving Jam in Synchronized Flow Spontaneous wide moving jam emergence in metastable synchronized flow is a phase transition from the synchronized flow phase to wide moving jam phase called an $S \rightarrow J$ transition. The $S \rightarrow J$ transition occurs due to the uninterrupted growth of a narrow moving jam that appears within the synchronized flow due to the $S \rightarrow J$ instability. The growing narrow moving jam can be considered a nucleus for the $S \rightarrow J$ transition.

Synchronized Flow Traffic Phase In three-phase traffic theory, the following definition (criterion) [S] of the synchronized flow phase (synchronized flow for short) in congested

traffic is made. The definition [S]: in contrast to the wide moving jam phase, the downstream front of the synchronized flow phase does *not* maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often *fixed* at a bottleneck. In other words, synchronized flow does not exhibit the characteristic jam feature [J].

Criteria [S] and [J] for Traffic Phases Criteria (definitions) [S] and [J] for traffic phases are spatiotemporal macroscopic criteria for traffic phases in congested traffic that define the synchronized flow and wide moving jam traffic phases, respectively.

Microscopic Criterion for Wide Moving Jam Within wide moving jams, there are regions in which traffic flow is interrupted: the inflow into the jam has no influence on the jam outflow. The flow interruption effect determines the microscopic criterion for the wide moving jam phase. If congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set of congested traffic are related to the synchronized flow phase.

Line J The line J is a characteristic line in the flow–density plane representing a steady propagation of the downstream front of a wide moving jam. The slope of the line J is determined by the mean velocity of the downstream jam front. All infinite numbers of states of traffic flow in the flow–density plane that lie on the line J are threshold states for wide moving jam existence and emergence. The line J separates all steady states of synchronized flow in the flow–density plane into two qualitatively different classes: (i) States below the line J are associated with homogeneous synchronized flow in which no wide moving jam can emerge spontaneously and persist. (ii) States on and above the line J are associated with metastable synchronized flow in which a wide moving jam can emerge and exist.

Synchronized Flow Patterns A synchronized flow pattern (SP) is a congested traffic pattern that consists of the synchronized flow phase only.

Pinch Effect in Synchronized Flow The pinch effect in synchronized flow is the effect of spontaneous occurrence and growth of narrow moving jams in synchronized flow (S→J instability). Within the associated pinch region of synchronized flow, a compression of synchronized flow is often observed.

General Congested Patterns A general pattern (GP) is a congested traffic pattern at a hypothetical isolated bottleneck that consists of both traffic phases in congested traffic – synchronized flow and wide moving jam. Firstly, synchronized flow occurs at the bottleneck. Due to the pinch effect in this synchronized flow, later and upstream of the bottleneck, narrow moving jams emerge spontaneously; jam amplitudes grow, while the jams propagate upstream within the pinch region. Some of the jams (or each of them) transform into wide moving jams at the upstream boundary of the pinch region of synchronized flow. Thus, GP structure consists of the pinch region of synchronized flow and a sequence of wide moving jams upstream.

Diagram of Congested Patterns The diagram of congested traffic patterns at a hypothetical isolated bottleneck shows regions of congested pattern existence and excitation depending on traffic demand and bottleneck characteristics.

Expanded Congested Patterns An expanded congested pattern (EP) is a congested traffic pattern whose synchronized flow affects two or more adjacent road bottlenecks.

Microscopic Criterion for Traffic Phases Within wide moving jams, there are regions in which traffic flow is interrupted: the inflow into the jam has no influence on the jam outflow. A sufficient condition for flow interruption determines the microscopic criterion for the wide moving jam phase. Based on this criterion, both the synchronized flow and wide moving jam phases can be identified in congested traffic from single-vehicle data measured even at one freeway location. This is because if congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase.

Microscopic Structure of Wide Moving

Jam The wide moving jam structure consists of alternations of regions in which traffic flow is interrupted and flow states of low speeds associated with moving blanks within the jam. The flow interruption intervals within the jam occur when some of the vehicles within the jam do not move.

Moving Blanks Within Wide Moving

Jam When vehicles meet a wide moving jam, they come to a stop at very different blank spaces (blanks for short) to each other. Later vehicles begin to move within the jam covering these blanks. Due to this low-speed vehicle motion, new blanks between vehicles occur upstream within the jam. A moving blank is a blank between vehicles within a wide moving jam, which moves upstream due to low-speed vehicle motion within the jam.

Heavy Bottleneck A heavy bottleneck is a bottleneck with a large bottleneck strength, i.e., a large bottleneck influence on traffic flow. The heavy bottleneck limits the average flow rate within a congested pattern occurring upstream of the bottleneck to very small values. A heavy bottleneck can occur due to, e.g., bad weather conditions, accidents, or heavy roadworks.

Traffic Congestion at Heavy Bottleneck Traffic congestion at a heavy bottleneck can be associated with the phenomena of random disappearance and appearance of the pinch region over time and a complex non-regular dynamics of wide moving jams upstream of the pinch region as well as with the phenomenon of the merger of wide moving jams into a mega-wide moving jam (mega-jam). When the bottleneck strength increases strongly, the pinch region disappears, and only the mega-jam survives, and synchronized flow remains only within its downstream front separating free flow and congested traffic.

spatiotemporal congested traffic patterns measured at a highway bottleneck, (ii) complex evolution of these congested patterns in space and time that occurs when traffic demand or/and bottleneck characteristics change, (iii) complex spatiotemporal phenomena associated with congested patterns occurring at two or more adjacent highway bottlenecks, (iv) transformations between various congested patterns that occur due to phase transitions between different traffic phases within the pattern, (v) diverse microscopic characteristics of traffic congestion associated with complex driver behavior within congested patterns, and (vi) complex non-regular dynamics of wide moving jams that can occur when a heavy bottleneck appears on a road.

Traffic congestion is a fact of life for many car drivers. For this reason, one of the aims of traffic and transportation research is to provide the understanding of highway traffic congestion that can be used for effective traffic management, traffic control, traffic organization, traffic assignment, and other engineering applications, which should increase traffic safety and result in high-quality mobility.

It should be noted that when the inflow rates onto a highway traffic network increase to very large values, it can turn out that congested traffic cannot be avoided in the network any more. Indeed, in many real traffic networks, there are not enough alternative routes to avoid traffic congestion at large enough traffic demand. Nevertheless, the application of intelligent transportation systems (ITS) can change characteristics of traffic congestion with the objective to increase traffic safety and comfort while moving in heavy congested traffic. It is assumed that the future vehicular traffic is mixed traffic flow consisting of human driving and automated driving (self-driving) vehicles. Therefore, the understanding of spatiotemporal features of heavy congested traffic in highway traffic is very important for the development of automated driving vehicles.

Definition of the Subject and Its Importance

Empirical observations show that traffic congestion exhibits extremely complex spatiotemporal features. Spatiotemporal features of traffic congestion are as follows: (i) diverse variety of

Introduction

There are a huge number of publications devoted to empirical and experimental investigations of

spatiotemporal features of vehicular traffic congestion (e.g., Edie and Foote (1958, 1960); Edie (1961); Edie et al. (1980); Koshi et al. (1983); Treiterer and Taylor (1966); Treiterer and Myers (1974); Treiterer (1975); Sugiyama et al. (2008); Jiang et al. (2015, 2017); Rehborn and Palmer (2008); Rehborn and Koller (2014); Rehborn et al. (2011a, b, 2017); Rempe et al. (2017, 2016); Molzahn et al. (2017) and references in Leutzbach (1988), May (1990), Haight (1963), Chowdhury et al. (2000), Helbing (2001), Nagatani (2002), Nagel et al. (2003), Kerner (2004b), Treiber and Kesting (2013), Elefteriadou (2014), Rehborn and Klenov (2009), and Hausken and Rehborn (2015)). In particular, one of the congested traffic patterns is a moving traffic jam (Edie and Foote 1958, 1960; Edie 1961; Edie et al. 1980; Koshi et al. 1983; Treiterer and Taylor 1966; Treiterer and Myers 1974; Treiterer 1975). A moving traffic jam (moving jam for short) is a localized structure of large vehicle density spatially limited by two jam fronts. Within the downstream jam front, vehicles accelerate escaping from the jam; within the upstream jam front, vehicles slow down approaching the jam. Both jam fronts (and, therefore, the jam as a whole structure) propagate upstream in traffic flow. Within the jam (i.e., between the jam fronts), vehicle density is large, and speed is very low (sometimes as low as zero). Moving jams have been studied empirically by many authors, in particular, in classic empirical works by Edie et al. (Edie and Foote 1958, 1960; Edie 1961; Edie et al. 1980), Treiterer et al. (Treiterer and Taylor 1966; Treiterer and Myers 1974; Treiterer 1975), and Koshi et al. (1983).

It is well-known that traffic congestion results from traffic breakdown that is a transition from free traffic flow (free flow for short) to congested traffic. Traffic breakdown occurs mostly at a bottleneck (Leutzbach 1988; May 1990; Daganzo 1997; Hall et al. 1992; Persaud et al. 1998; Hall and Agyemang-Duah 1991; Elefteriadou et al. 1995; Elefteriadou 2014). Two main types of bottlenecks can be distinguished: (i) a road bottleneck and (ii) a moving bottleneck. The road bottleneck is fixed at some road location. The moving

bottleneck is caused by a slow-moving vehicle in traffic flow. Road bottlenecks are caused for example by on- and off-ramps, road gradients, road works, a decrease in the number of road lines (in the flow direction), traffic signal, etc. (e.g., Leutzbach 1988; May 1990; Daganzo 1997).

The most important empirical feature of traffic breakdown in a network is the nucleation nature of traffic breakdown found in real-field traffic data. The term *nucleation nature* of traffic breakdown means that traffic breakdown occurs in a metastable free flow (Kerner 1998a, b, c, 1999a, b, c). The metastability of free flow is as follows. There can be many speed (density, flow rate) disturbances in free flow. Amplitudes of the disturbances can be very different. When a disturbance occurs randomly whose amplitude is larger than a critical one, then traffic breakdown occurs. Such a disturbance resulting in the breakdown is called a *nucleus* for traffic breakdown. Otherwise, if the disturbance amplitude is smaller than the critical one, the disturbance decays; as a result, no traffic breakdown occurs. In empirical observations, traffic breakdown at a highway bottleneck is a phase transition from metastable free flow to congested traffic whose downstream front is usually fixed at the bottleneck. Such congested traffic is called synchronized flow. Therefore, empirical traffic breakdown is a phase transition from metastable free flow to synchronized flow ($F \rightarrow S$ transition):

- The nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at a network bottleneck is an empirical fundamental of transportation science.

Unfortunately, generally accepted classical traffic and transportation theories, which have had a great impact on the understanding of many empirical traffic phenomena, have nevertheless failed by their applications in the real world. Even several decades of a very intensive effort to improve and validate network optimization and control models based on the classical traffic and transportation theories have had no success. Indeed, there can be found no examples where online implementations of the network

optimization models based on these classical traffic and transportation theories could reduce congestion in real traffic and transportation networks. In the book Kerner (2017a), we have shown that this failure of classical traffic and transportation theories can be explained as follows: The classical traffic and transportation theories are not consistent with the nucleation nature of traffic breakdown. Consequently, as has already been explained in reviews (Kerner 2013b, 2015b, 2016) and the book (Kerner 2017a) as well as in this Encyclopedia (Kerner 2017b, 2018a), earlier traffic flow theories and models reviewed, for example, in Leutzbach (1988), May (1990), Haight (1963), Daganzo (1997), Wiedemann (1974), Whitham (1974), Cremer (1979), Gartner et al. (1997), Wolf (1999), Chowdhury et al. (2000), Helbing (2001), Nagatani (2002), Nagel et al. (2003), Mahnke et al. (2005), Bellomo et al. (2002), Treiber and Kesting (2013), Elefteriadou (2014), Mahmassani (2005), Schreckenberg and Wolf (1998), Helbing et al. (2000), Schadschneider et al. (2007), Fukui et al. (2003), and Hoogendoorn et al. (2005), cannot explain and predict the nucleation nature of empirical traffic breakdown ($F \rightarrow S$ transition). These traffic flow theories and models (see, e.g., reviews Leutzbach (1988), May (1990), Haight (1963), Daganzo (1997), Wiedemann (1974), Whitham (1974), Cremer (1979), Gartner et al. (1997), Wolf (1999), Chowdhury et al. (2000), Helbing (2001), Nagatani (2002), Nagel et al. (2003), Mahnke et al. (2005), Bellomo et al. (2002), Treiber and Kesting (2013), and Elefteriadou (2014)) are the theoretical basis for the reliability analysis of applications of automated driving (self-driving) vehicles as well as other intelligent transportation systems (ITS) in vehicular traffic. Therefore, many results of the analysis of ITS applications made with the use of the classical traffic flow models are invalid for real traffic (this criticism has been considered in more details in the book Kerner (2017a) and the review articles Kerner (2017b) and Kerner (2018a) of this Encyclopedia).

The nucleation nature of empirical traffic breakdown was understood only during last 20 years. In contrast, the generally accepted

fundamentals and methodologies of traffic and transportation theory were introduced in the 1950s–1960s. Thus, the scientists whose ideas led to the classical traffic and transportation theories could not know the empirical nucleation nature of traffic breakdown. Respectively, in 1996–2002, the author introduced and developed the three-phase traffic theory (three-phase theory for short) (Kerner 1998a, b, c, 1999a, b, c, 2004b).

In the three-phase theory, besides the free flow traffic phase, there are two traffic phases in congested traffic: synchronized flow and wide moving jam. Thus, there are three traffic phases in this theory:

1. Free flow (F).
2. Synchronized flow (S).
3. Wide moving jam (J).

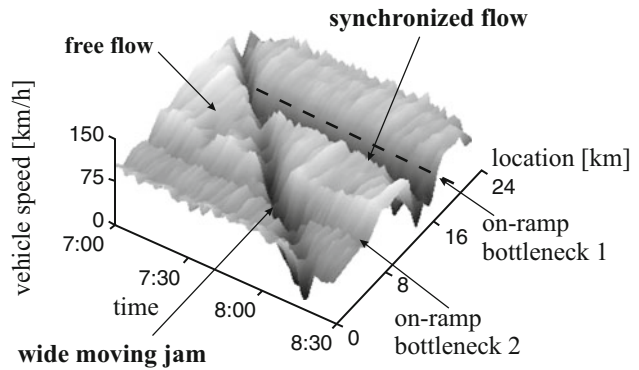
The wide moving jam and synchronized flow traffic phases in congested traffic are defined through spatiotemporal empirical criteria [J] and [S]. The definition of the wide moving jam phase [J]: a wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks.

The definition of the synchronized flow phase [S]: in contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

Note that the downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow, vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow downstream of the synchronized flow. To apply the traffic phase definitions [J] and [S], spatiotemporal features of congested traffic should be known. The criteria [J] and [S] also mean that if a congested traffic state is not related to the wide moving jam phase, then with certainty the state is associated with the synchronized flow phase. This is because in the three-phase theory, congested traffic can be either

Spatiotemporal Features of Traffic Congestion,

Fig. 1 Three traffic phases in real empirical traffic data measured through road detectors on highway A5-South in Germany: average vehicle speed in space and time. (Adapted from Kerner 2004b)



within the synchronized flow phase or within the wide moving jam phase. In other words, if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase. For example, the definition [S] means that moving jams, which are caught at a bottleneck, are related to the synchronized flow phase rather than to the wide moving jam phase.

The traffic phase definitions [J] and [S] are illustrated by measured traffic data shown in Fig. 1. There are two qualitatively different types of congested patterns. The first congested pattern propagates through both on-ramp bottlenecks (labeled “on-ramp bottleneck 1” and “on-ramp bottleneck 2” in Fig. 1) while maintaining the mean velocity of the downstream front of the pattern. In accordance with the definition [J], this congested pattern is associated with the wide moving jam traffic phase in congested traffic. The downstream fronts of another congested pattern is fixed at on-ramp bottleneck 1. In accordance with the definition [S], these congested patterns are associated with the synchronized flow traffic phase.

It must be noted that some of the classical traffic flow models reviewed, for example, in Leutzbach (1988), May (1990), Haight (1963), Daganzo (1997), Wiedemann (1974), Whitham (1974), Cremer (1979), Gartner et al. (1997), Wolf (1999), Chowdhury et al. (2000), Helbing (2001), Nagatani (2002), Nagel et al. (2003),

Mahnke et al. (2005), Bellomo et al. (2002), Treiber and Kesting (2013), and Elefteriadou (2014), can explain and show many of the characteristic features of moving jams. However, none of the classical traffic flow models can explain and show the metastability of free flow with respect to an $F \rightarrow S$ transition (traffic breakdown). The models (Leutzbach 1988; May 1990; Haight 1963; Daganzo 1997; Wiedemann 1974; Whitham 1974; Cremer 1979; Gartner et al. 1997; Wolf 1999; Chowdhury et al. 2000; Helbing 2001; Nagatani 2002; Nagel et al. 2003; Mahnke et al. 2005; Bellomo et al. 2002; Treiber and Kesting 2013; Elefteriadou 2014) cannot also show and explain many nonlinear features of synchronized flow. For this reason, in this article, when simulation results are presented for the illustration of empirical data, the simulation results have been made with the use of microscopic traffic flow models in the framework of the three-phase theory.

In the articles Kerner (2017b, 2018a) of this Encyclopedia, the nucleation empirical feature of traffic breakdown and a critical discussion of earlier traffic flow modeling approaches to traffic breakdown have already been reviewed. In Kerner (2018b, 2018c), this critical analysis has been applied for a study of the effect of automated driving (autonomous driving) in the framework of the three-phase theory on features of mixed traffic flow consisting of automated driving vehicles and human driving vehicles. The main focus of this article is a review of empirical *spatiotemporal features* of

traffic congested patterns resulting from traffic breakdown as well as an explanation of these empirical traffic patterns.

The article is organized as follows. Firstly, in section “[Empirical Phase Transitions in Traffic Flow](#),” we make an overview of empirical features of phase transitions in traffic flow that determine the spatiotemporal complexity of congested traffic. In this section, we make also a discussion of two qualitatively different instabilities of synchronized flow in the three-phase theory that are responsible for the features of the phase transitions in traffic flow. In section “[Empirical Characteristic Parameters of Wide Moving Jam Propagation](#),” we consider characteristic parameters of wide moving jam propagation, which are needed for the understanding of spatiotemporal traffic congested patterns occurring on highways. In section “[Microscopic Features of Traffic Phases in Congested Traffic](#),” we consider some microscopic features of traffic congestion. Spatiotemporal features of moving jam emergence in synchronized flow are studied in section “[Moving Jam Emergence in Framework of Three-Phase Traffic Theory](#).” Spatiotemporal features of congested patterns occurring at highway bottlenecks, a diagram of congested patterns, congested pattern interaction at adjacent bottlenecks, as well as reproducible and predictable features of spatiotemporal congested traffic patterns are discussed in section “[Congested Patterns at Road Bottlenecks](#).” In section “[Congested Patterns at Heavy Bottlenecks](#),” we discuss features of congested patterns at heavy bottlenecks caused, for example, by bad weather conditions, accidents, or heavy roadworks. In section “[“Jam-Absorption” Effect in Highway and City Traffic](#),” we discuss a “jam-absorption” effect, i.e., the effect of moving jam dissolution occurring in congested traffic through strong driver speed adaptation; due to jam dissolution, congested traffic consists only of the synchronized flow phase. Finally, in section “[Conclusions: Fundamental Empirical Features of Spatiotemporal Congested Traffic Patterns](#),” we make a summary of fundamental empirical spatiotemporal features of traffic congested patterns.

Empirical Phase Transitions in Traffic Flow

Spatiotemporal features of congested traffic depend considerably on characteristics of phase transitions that occur between three traffic phases F (free flow), S (synchronized flow), and J (wide moving jam). For this reason, in this section, we discuss empirical phase transitions that can be observed in real highway traffic.

The definition of the traffic phases made in the three-phase theory (section “[Introduction](#)”) leads to the explanation of the nature of phase transitions observed in real-field traffic data (Kerner 1998a, b, c, 1999a, b, c, 2000a, b, 2001, 2002a, b). Based on a study of real-field traffic data, it has been found that both a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) and a phase transition from synchronized flow to a wide moving jam (called $S \rightarrow J$ transition) exhibit the nucleation nature (Kerner 1998a, b, c, 1999a, b, c).

The empirical nucleation nature of the $F \rightarrow S$ transition is as follows (Kerner 1998a, b, c, 1999a, b, c, 2004b):

- The phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) occurs in a metastable state of free flow with respect to the $F \rightarrow S$ transition at a highway bottleneck.

To understand the sense of the term *metastable state of free flow with respect to the $F \rightarrow S$ transition at the bottleneck*, we consider a local speed disturbance in free flow that causes a local speed decrease at the bottleneck. The disturbance is localized at the bottleneck. We use the term “the amplitude of the disturbance”: The larger the disturbance amplitude, the smaller the free flow speed within the disturbance in comparison with the free flow speed outside the disturbance. If the amplitude of the disturbance is small enough, then no $F \rightarrow S$ transition occurs at the bottleneck. However, when the amplitude of a local speed disturbance in the metastable free flow is equal to or exceeds a critical amplitude, the $F \rightarrow S$ transition does occur. A local speed disturbance occurring in the metastable free flow that leads

to the $F \rightarrow S$ transition is called a nucleus for the $F \rightarrow S$ transition. Respectively, a local speed disturbance in the metastable free flow with a critical amplitude required for traffic breakdown ($F \rightarrow S$ transition) is called a critical nucleus. In other words, in real traffic, there is the metastability of free flow with respect to the $F \rightarrow S$ transition at the bottleneck. As above mentioned, the term *traffic breakdown at a highway bottleneck* is a synonym of the term *$F \rightarrow S$ transition at the bottleneck*. Therefore, the term *a nucleus for $F \rightarrow S$ transition at the bottleneck* is a synonym of the term *a nucleus for traffic breakdown at the bottleneck*. Detailed empirical and theoretical studies of the nucleation nature of traffic breakdown ($F \rightarrow S$ transition) at highway bottlenecks have been made in the book Kerner (2017a) as well as in the article Kerner (2018a) of this Encyclopedia.

The empirical nucleation nature of the $S \rightarrow J$ transition is as follows (Kerner 1998a, b, c, 1999a, b, c, 2004b):

- The phase transition from synchronized flow to a wide moving jam(s) ($S \rightarrow J$ transition) occurs in a metastable state of synchronized flow with respect to the $S \rightarrow J$ transition.

We consider a local speed disturbance in synchronized flow that causes a local speed decrease. The disturbance propagates upstream in synchronized flow. When the amplitude of the local speed disturbance in this metastable synchronized flow is small enough, the local disturbance decays. However, when the amplitude of a local speed disturbance in the metastable synchronized flow is equal to or exceeds a critical amplitude, the $S \rightarrow J$ transition does occur. A local speed disturbance occurring in the metastable synchronized flow that leads to the $S \rightarrow J$ transition is called a nucleus for the $S \rightarrow J$ transition.

These empirical results (Figs. 2 and 3) lead to the following conclusion of the three-phase theory (Kerner 1998c):

- Wide moving jams emerge spontaneously in free flow due to a sequence of two phase transitions: Firstly, the $F \rightarrow S$ transition occurs at

the bottleneck leading to synchronized flow upstream of the bottleneck. Synchronized flow propagates upstream of the bottleneck. Within the emergent synchronized flow, at some distance upstream of the bottleneck, wide moving jams emerge spontaneously ($S \rightarrow J$ transition) (Figs. 2b and 3). This sequence of the phase transitions is called the $F \rightarrow S \rightarrow J$ transitions.

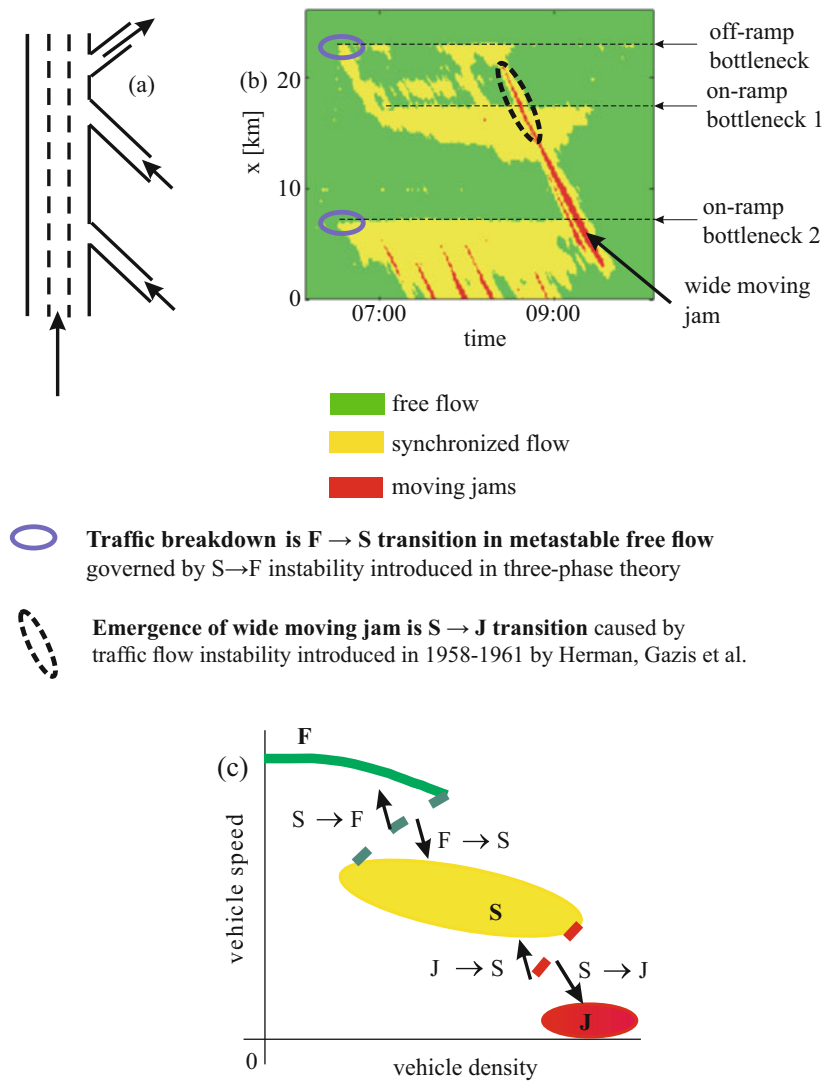
Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow

Both the $F \rightarrow S$ transition (traffic breakdown) and the $S \rightarrow J$ transition that occur in space and time in real traffic (Figs. 2b and 3) can also be presented in the speed–density, speed–flow rate, and flow–density planes by double Z characteristics for phase transitions (called 2Z characteristic for phase transitions) (Figs. 2c and 4) (Kerner 2004b; Kerner and Klenov 2003). Obviously, there are also return $S \rightarrow F$ and $J \rightarrow S$ transitions between the three traffic phases (Figs. 3 and 4).

Arrows $F \rightarrow S$ and $S \rightarrow F$ shown in Fig. 4 mark, respectively, the $F \rightarrow S$ transition and the $S \rightarrow F$ transition in the speed–density, speed–flow rate, and flow–density planes associated with real $F \rightarrow S$ and $S \rightarrow F$ transitions occurring in space and time (Fig. 3). Within wide moving jams, low speed states are observed that are probably associated with low speed within wide moving jams associated with the so-called moving blanks (labeled by “low speed states within wide moving jam” in Fig. 4) that will be discussed in section “Microscopic Features of Traffic Phases in Congested Traffic.”

We see that there are many diverse transitions between the three traffic phases as well as between diverse traffic states within the phases associated with congested pattern emergence and dissolution. The phase transitions and transitions within the same traffic phase could exhibit a variety of hysteresis effects (Fig. 4). The phase transitions in traffic and transitions within the same traffic phase cannot be distinguished from each other without knowledge of the whole *spatiotemporal dynamics* of the congested pattern (Fig. 3).

As discussed in section “Introduction,” for $F \rightarrow S$ transition occurrence, a local speed

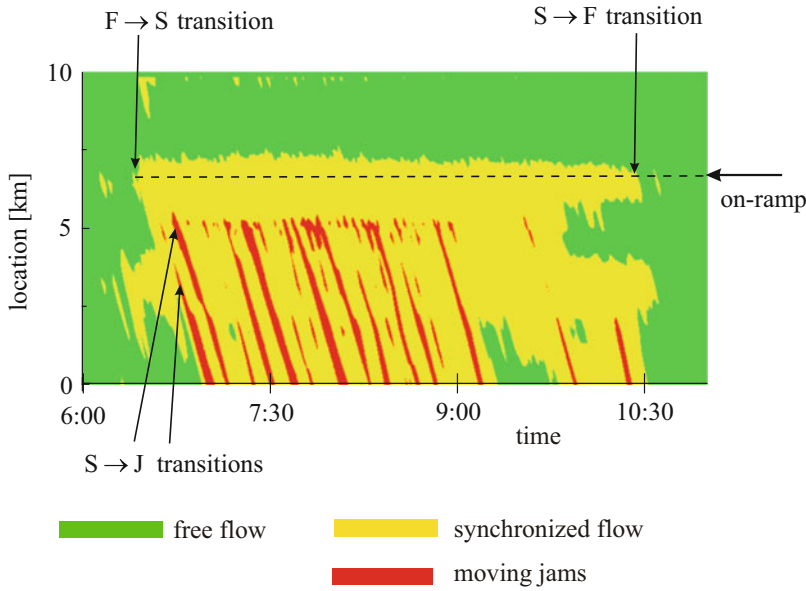


Spatiotemporal Features of Traffic Congestion, Fig. 2 Empirical example of phase transitions: (a) Sketch of section of three-lane highway in Germany with three bottlenecks. (b) Speed data measured on April 20, 1998, on freeway A5-South in Germany with road detectors installed along road section in (a); data is presented in space and time with averaging method described in Sect.

C.2 of Kerner et al. (2013b). (c) Hypothesis of three-phase theory about features of phase transitions in traffic flow: qualitative 2Z-characteristic for phase transitions. F, free flow phase; S, synchronized flow phase; J, wide moving jam phase. (Adapted from Kerner 2004b; Kerner and Klenov 2003)

disturbance should occur in free flow at the bottleneck whose amplitude is equal or larger than a critical disturbance amplitude. The larger the density (flow rate) downstream of the bottleneck, the smaller the critical disturbance amplitude.

Respectively, for $S \rightarrow J$ transition occurrence, a local speed disturbance should appear in synchronized flow upstream of the bottleneck whose amplitude is equal or larger than another critical disturbance amplitude. For this reason, in a



Spatiotemporal Features of Traffic Congestion,

Fig. 3 Empirical example of $F \rightarrow S \rightarrow J$ transitions in real traffic: speed in space and time within a congested pattern occurring at on-ramp bottleneck (this bottleneck is on-ramp bottleneck 2 in Fig. 2a); 1-min averaged data measured on March 23, 1998, on freeway A5-South in Germany with road detectors installed along three-lane road section; data is presented in space and time with

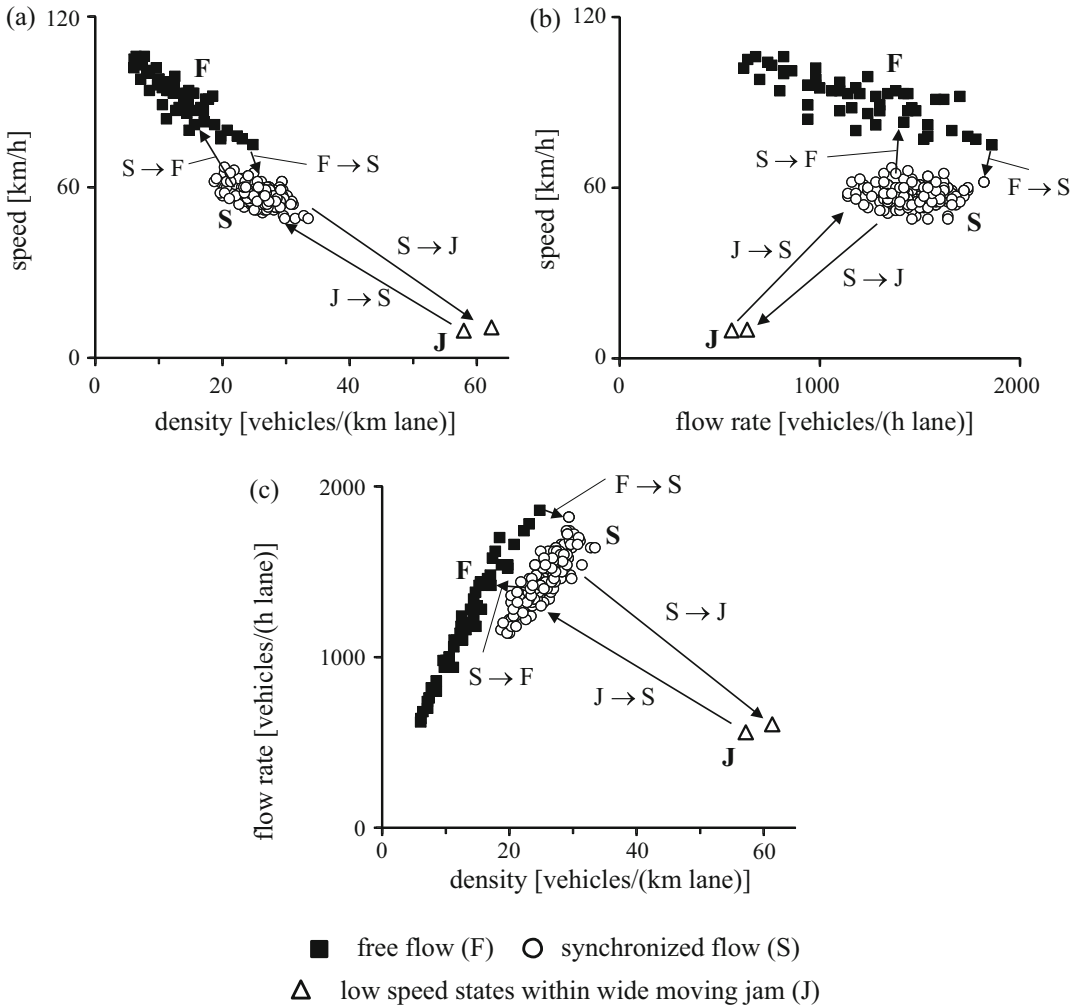
averaging method described in Sect. C.2 of Kerner et al. (2013b). Arrows $F \rightarrow S$ and $S \rightarrow F$ mark symbolically, respectively, the $F \rightarrow S$ and $S \rightarrow F$ transitions at the location of on-ramp bottleneck, arrows $S \rightarrow J$ mark symbolically $S \rightarrow J$ transitions related to the emergence of two first wide moving jams within synchronized flow. (Adapted from Kerner 1998a, b, c)

theoretical 2Z-characteristic (Fig. 2c) between states of free flow, synchronized flow, and low speed states within wide moving jams, there are states (shown by dashed curves in Fig. 2c) associated with average speeds within the critical disturbances (critical nuclei) required for phase transition occurrence between the related traffic phases.

It should be noted that in addition to the critical disturbance amplitude, the critical disturbance (critical nucleus) is also characterized by a *critical spatial shape*, i.e., by the critical speed (density) spatial distribution within the disturbance. This conclusion of the three-phase theory is valid for both $F \rightarrow S$ and $S \rightarrow J$ transitions. However, for simplicity we neglect the critical spatial shape of the disturbances in Fig. 2c, in which critical disturbances (critical nuclei for the phase transitions) are qualitatively presented by dashed curves

between states of the free flow phase (F) and the synchronized flow phase (S) associated with the average speed within the critical nuclei for $F \rightarrow S$ transitions and between states of the synchronized flow phase (S) and the wide moving jam phase (J) associated with the average speed within the critical nuclei for $S \rightarrow J$ transitions.

A theoretical 2Z-characteristic for traffic breakdown derived through simulations of the Kerner-Klenov microscopic stochastic three-phase model (see Fig. 37 of a review article Kerner (2018a) in this Encyclopedia) confirms the nucleation nature of the phase transitions in traffic flow discussed above. It should be noted that a detailed consideration of simulation results of features of the phase transitions between the traffic phases can be found in the books Kerner (2004b, 2017a) as well as in the review articles Kerner (2009a, 2018a).



Spatiotemporal Features of Traffic Congestion, Fig. 4 Empirical 2Z characteristics for phase transitions: Presentation of phase transitions shown in Fig. 3 in the speed–density (a), speed–flow rate (b) and flow–density

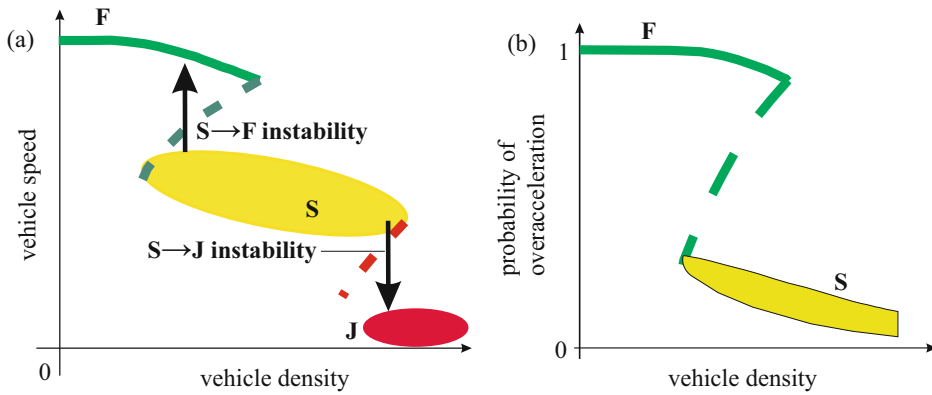
planes (c) (points J are related to the second of the wide moving jams in Fig. 3). (Adapted from Kerner 2004b; Kerner and Klenov 2003)

Two Qualitatively Different Instabilities of Synchronized Flow in Three-Phase Theory

In the three-phase theory (Kerner 1998a, b, c, 1999a, b, c, 2004b, 2009c, 2017a), the empirical $F \rightarrow S$, $S \rightarrow F$, and $S \rightarrow J$ phase transitions discussed in sections “Empirical Phase Transitions in Traffic Flow” and “Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow” (Figs. 2, 3, and 4) have been explained through the hypothesis about the existence of two qualitatively

different instabilities in synchronized flow: an $S \rightarrow F$ instability (labeled by “ $S \rightarrow F$ instability” in Fig. 5a) and $S \rightarrow J$ instability (labeled by “ $S \rightarrow J$ instability” in Fig. 5a).

1. The $S \rightarrow F$ instability occurs due to a discontinuous character of over-acceleration probability (Fig. 5b) associated with a time delay in driver over-acceleration. The $S \rightarrow F$ instability leads to a growing wave of the local **increase** in



Spatiotemporal Features of Traffic Congestion, Fig. 5 Qualitative illustration of two qualitatively different instabilities in synchronized flow in the three-phase theory: (a) Qualitative 2Z-characteristic for phase

transitions. (b) Qualitative discontinuous character of over-acceleration probability. (Adapted from Kerner 2004b, 2009c, 2017a)

the speed in synchronized flow. The $S \rightarrow F$ instability has been introduced in the three-phase theory (Kerner 2004b, 2009c, 2017a, 2015a). The $S \rightarrow F$ instability governs the nucleation nature of traffic breakdown ($F \rightarrow S$ transition). The $S \rightarrow F$ instability that has been considered in the book Kerner (2017a) will not be discussed in this review article.

2. The $S \rightarrow J$ instability leads to a growing wave of the local **decrease** in the speed in synchronized flow. The $S \rightarrow J$ instability occurs due to the driver over-deceleration effect (driver reaction time). The driver over-deceleration effect is the reason for the classical traffic flow instability of Herman, Gazis, Montroll, Potts, Rothery, and Chandler (Gazis et al. 1961, 1959; Chandler et al. 1958; Herman et al. 1959). Therefore, the $S \rightarrow J$ instability is associated with the classical traffic flow instability (Gazis et al. 1961, 1959; Chandler et al. 1958; Herman et al. 1959). The $S \rightarrow J$ instability can result in an $S \rightarrow J$ transition; however, the $S \rightarrow J$ instability should not necessarily lead to the $S \rightarrow J$ transition (see detailed explanations in section “Moving Jam Emergence in Framework of Three-Phase Traffic Theory” below).

Sequence of $F \rightarrow S \rightarrow J$ Transitions

Wide moving jams do not emerge spontaneously in free flow: no spontaneous phase transition from

the free flow phase to the wide moving jam phase ($F \rightarrow J$ transition for short) has been observed (Kerner 2004b). Wide moving jams can emerge *spontaneously only* in the synchronized flow phase ($S \rightarrow J$ transition).

In synchronized flow at higher speeds, wide moving jams should not necessarily emerge spontaneously. Observations show that the larger the density in synchronized flow, the more likely is spontaneous moving jam emergence in that synchronized flow. Thus, a wide moving jam emerges in an initial free flow due to a sequence of two phase transitions: Firstly, an $F \rightarrow S$ transition occurs, and synchronized flow emerges. Later and usually at other freeway locations than the location of the $F \rightarrow S$ transition, an $S \rightarrow J$ transition occurs spontaneously leading to wide moving jam emergence. As mentioned in section “Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow,” this sequence of phase transitions is called the $F \rightarrow S \rightarrow J$ transitions.

As explained in section “Two Qualitatively Different Instabilities of Synchronized Flow in Three-Phase Theory,” an $S \rightarrow J$ transition results from the $S \rightarrow J$ instability. Therefore, conditions at which the $S \rightarrow J$ instability occurs in synchronized flow are of a great importance for the understanding of spatiotemporal features of congested traffic. The conditions and features of the $S \rightarrow J$ instability will be considered in section “Moving

[Jam Emergence in Framework of Three-Phase Traffic Theory.](#)” However, to understand results of this section, firstly, we should discuss the line J and characteristic parameters of wide moving jam propagation in free flow revealed in 1994 (Kerner and Konhäuser 1994) (section [“Empirical Characteristic Parameters of Wide Moving Jam Propagation”](#)) as well as microscopic features of narrow and wide moving jams (section [“Microscopic Features of Traffic Phases in Congested Traffic.”](#))

Empirical Characteristic Parameters of Wide Moving Jam Propagation

The empirical feature of a steady propagation of the downstream front of wide moving jams is shown in Fig. 6 (Kerner and Rehborn 1996a, b, 1997). It can be seen that a sequence of two moving jams propagates through different states of traffic flow and through a bottleneck while maintaining the downstream jam front velocity. Therefore, these moving jams belong to the wide moving jam phase. In contrast, there is a congested traffic in which speed is much lower than in free flow (compare vehicle speeds in Fig. 6c, d). The downstream front of this congested traffic flow, where vehicles accelerate to free flow, is fixed at the bottleneck (dashed line in Fig. 6a). Therefore, this congested traffic belongs to the synchronized flow phase.

We can see that whereas in wide moving jams both the speed and flow rate are very low (sometimes as low as zero), in synchronized flow the flow rate is high (compare the flow rates within the wide moving jams and within synchronized flow in Fig. 6d). The vehicle speed in synchronized flow is considerably lower than in free flow. However, the flow rates in the free flow and synchronized flow phases can be close to one another (Fig. 6c, d).

The critical conclusion about earlier traffic flow models made in section [“Introduction”](#) does not concern characteristic parameters of wide moving jam propagation, which was firstly discovered by Kerner and Konhäuser in 1994 and

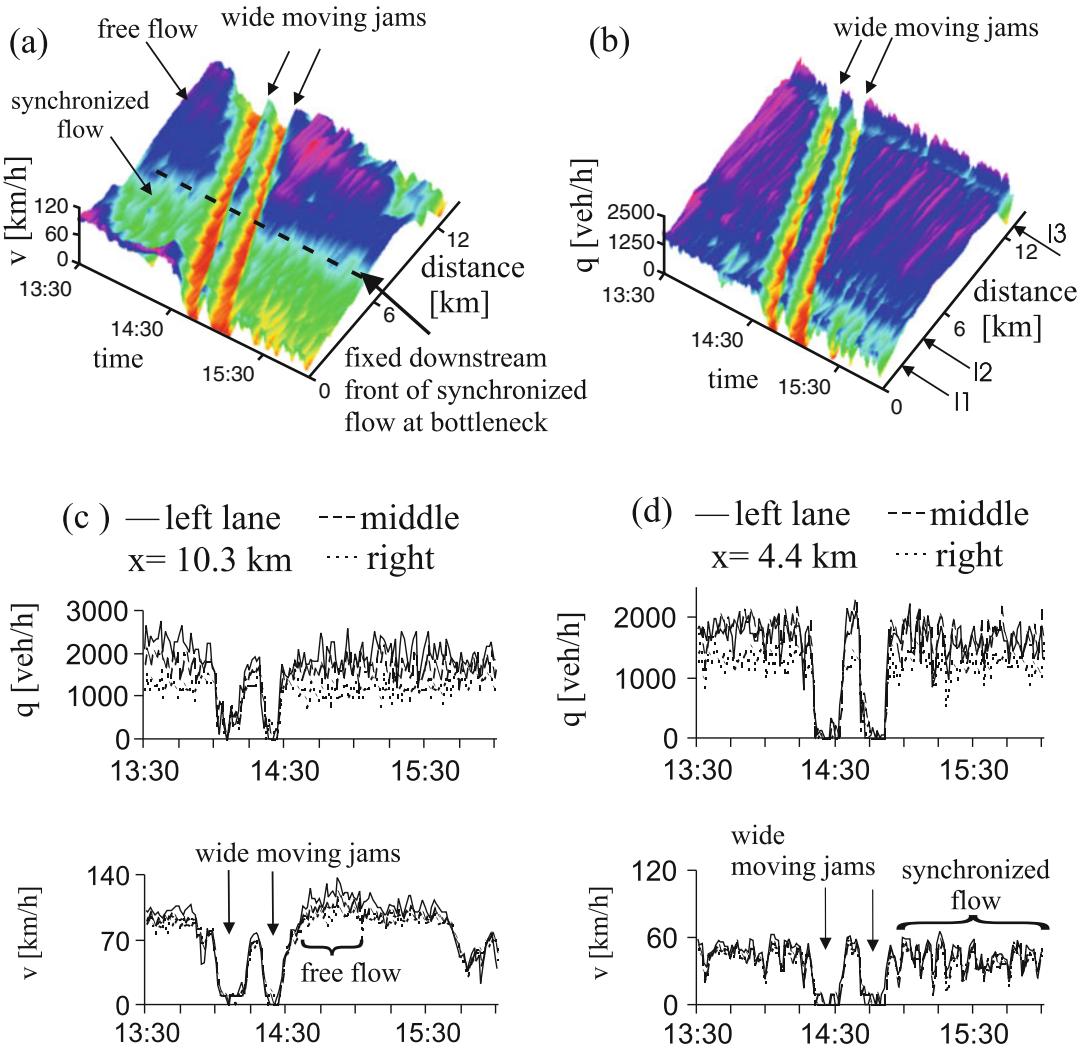
later incorporated in a number of different traffic flow models (see, e.g., references in Helbing 2001 and Treiber and Kesting 2013). The Kerner-Konhäuser theory of wide moving jam propagation has been confirmed by empirical data in which wide moving jam propagation has been found (Fig. 6) (Kerner and Rehborn 1996a, b, 1997).

The characteristic feature of wide moving jam propagation $[J]$, which defines the wide moving jam phase in congested traffic, i.e., a constancy of the mean velocity v_g of the downstream front of the jam while the jam propagates through bottlenecks, can be presented by a line in the flow–density plane. This line is called the line J (Figs. 7 and 8).

The slope of the line J is equal to the characteristic mean velocity v_g of the downstream jam front. If in the wide moving jam outflow free flow is formed, then the flow rate q_{out} in this jam outflow and the related density ρ_{min} give the left coordinates of the line J ; q_{out} , ρ_{min} , and the related mean vehicle speed v_{max} are characteristic jam parameters (Fig. 9). When control parameter of traffic (weather, percentage of long vehicles, etc.) do not change, the jam characteristic parameters do not depend on initial conditions, and they are the same for different wide moving jams. The right coordinates of the line J are related to the traffic variables within the jam, the density ρ_{max} , and the average vehicle speed v_{min} . For simplicity we assume here that $v_{min} = 0$ (low speed states within wide moving jams associated with the so-called moving blanks have been studied in Kerner et al. (2006a, b), and Kerner (2009a); see section [“Microscopic Features of Traffic Phases in Congested Traffic”](#)).

It has been predicted in Kerner and Konhäuser (1994) and proven empirically in Kerner and Rehborn (1996a, b) that all states of free flow that are on and above the line J (the flow rates q in free flow that satisfy condition $q \geq q_{out}$ in Figs. 7 and 8) are metastable states of free flow with respect to an $F \rightarrow J$ transition (wide moving jam emergence in free flow). However, as proven empirically (Kerner 1998c, 2004b, 2017a), the

A5-North, October 9, 1992

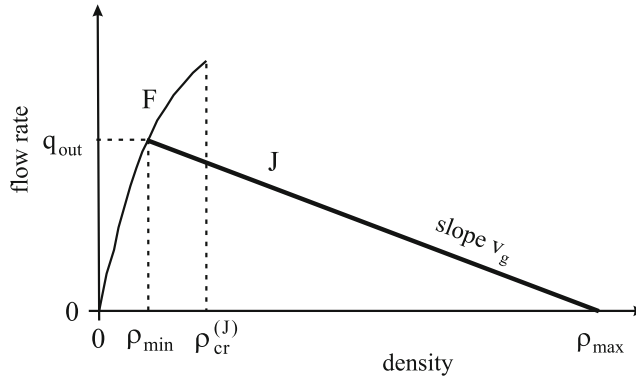


Spatiotemporal Features of Traffic Congestion, Fig. 6 Empirical characteristic parameters of wide moving jam propagation. (a, b) Vehicle speed (a) and flow rate averaged across all freeway lanes (b) as functions of time and space. (c, d) Flow rate and average vehicle speed at

two different freeway locations (c) and (d) in different freeway lanes. In (b), three intersections with other freeways are labeled by I1-I3 that show approximate locations of possible freeway bottlenecks. (Adapted from Kerner and Rehborn 1996a, b, 1997)

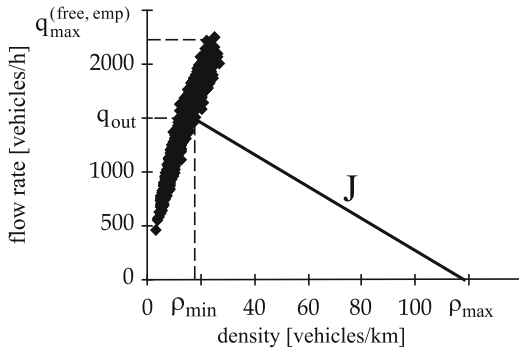
$F \rightarrow J$ transition does not occur *spontaneously* in real free flow. Rather than the $F \rightarrow J$ transition, an $F \rightarrow S$ transition determines the spontaneous occurrence of traffic breakdown at a bottleneck in all empirical traffic data. This is because in free flow, which is metastable free flow both with respect to the $F \rightarrow J$ transition and with respect to

the $F \rightarrow S$ transition, a nucleus required for the $F \rightarrow J$ transition is considerably larger than a nucleus required for the $F \rightarrow S$ transition. This explains why a spontaneous $F \rightarrow J$ transition has not been observed in real free flow (for more detailed explanations, see Chap. 8 of the book Kerner 2017a).



Spatiotemporal Features of Traffic Congestion, Fig. 7 Qualitative representation of the characteristic line for the downstream front of a wide moving jam (line J) whose slope is equal to the downstream jam front

velocity v_g , and explanation of the metastability of states of free flow (curve F) with respect to moving jam formation. (Adapted from Kerner and Konhäuser 1994)



Spatiotemporal Features of Traffic Congestion, Fig. 8 Empirical line J associated with Fig. 6 that represents the steady propagation of the downstream front of the downstream jam in the sequence of two empirical wide moving jams in the flow-density plane. Black diamonds show free flow data averaged across the road. (Adapted from Kerner and Rehborn 1996a, b, 1997)

Microscopic Features of Traffic Phases in Congested Traffic

Microscopic Criterion for Traffic Phases in Congested Traffic

The spatiotemporal criteria for a wide moving jam $[J]$, which distinguish the wide moving jam and synchronized flow phases, can be explained by a traffic flow discontinuity within a wide moving jam. This means that traffic flow is interrupted by the wide moving jam: there is no influence of the inflow into the jam on the jam outflow (Kerner

2004b). A difference between the jam inflow and the jam outflow changes the jam width (in the longitudinal direction) only. This *traffic flow interruption effect* is a general effect for each wide moving jam.

The jam outflow becomes independent of the jam inflow, when the traffic flow interruption effect within a moving jam occurs. In a hypothetical case, when all vehicles within a moving jam do not move, the criterion for the traffic flow interruption effect within the jam is

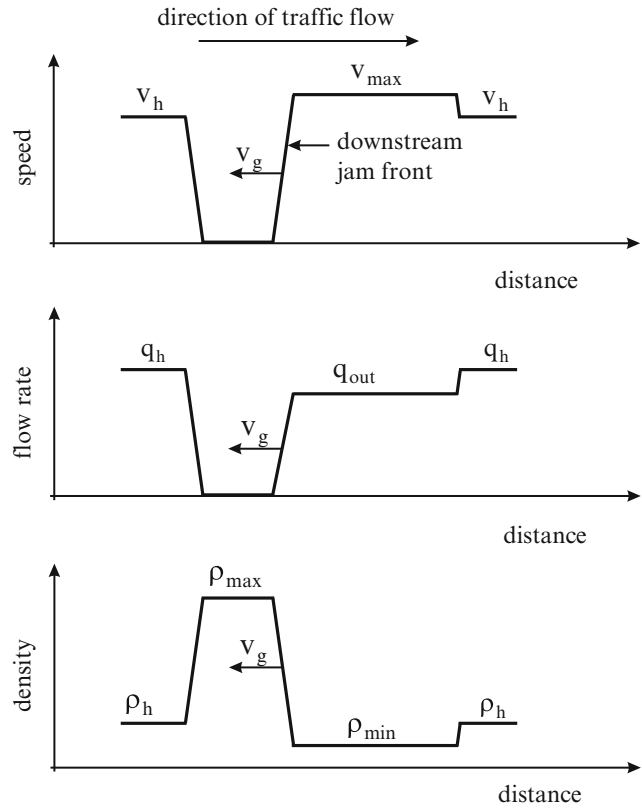
$$I = \frac{\tau_J}{\tau_{\text{del}}^{(\text{acc})}(0)} \gg 1, \quad (1)$$

where τ_J is the jam duration, i.e., the time interval between the upstream and downstream jam fronts passing a road location, and $\tau_{\text{del}}^{(\text{acc})}(0)$ is the mean time delay in vehicle acceleration at the downstream jam front from a vehicle standstill, $\tau_{\text{del}}^{(\text{acc})}(0)$ determines the jam outflow rate (Kerner 2004b), and I is approximately equal to the vehicle number stopped within the jam. Note that corresponding to empirical results $\tau_{\text{del}}^{(\text{acc})}(0) \approx 1.5 - 2$ s (Kerner 2004b).

In real traffic, vehicles come to a stop at very different blank spaces to each other when they meet a wide moving jam. We call these blank spaces as *blanks* within the jam. Later the vehicles begin to cover these blanks within the jam. This vehicle motion within the jam occurs at low

Spatiotemporal Features of Traffic Congestion,

Fig. 9 Characteristic parameters of wide moving jams. Schematic representation of a wide moving jam at a fixed time. Spatial distribution of vehicle speed v , flow rate q , and vehicle density ρ in the wide moving jam, which propagates through a homogeneous state of free flow with speed v_h , flow rate q_h , and density ρ_h . (Adapted from Kerner and Konhäuser 1994)



vehicle speeds, and it creates new blanks upstream. Thus, low speed states that cover blanks within the jam are associated with upstream moving of the blanks within the jam (moving blanks for short; for more details see Kerner et al. 2006b). In addition to (1), to study flow interruption within a wide moving jam in empirical single-vehicle data measured at a road location (Figs. 10 and 11), we consider a *sufficient* criterion for flow interruption within the jam that is as follows (Kerner et al. 2006a, 2007, 2006b; Kerner 2009c):

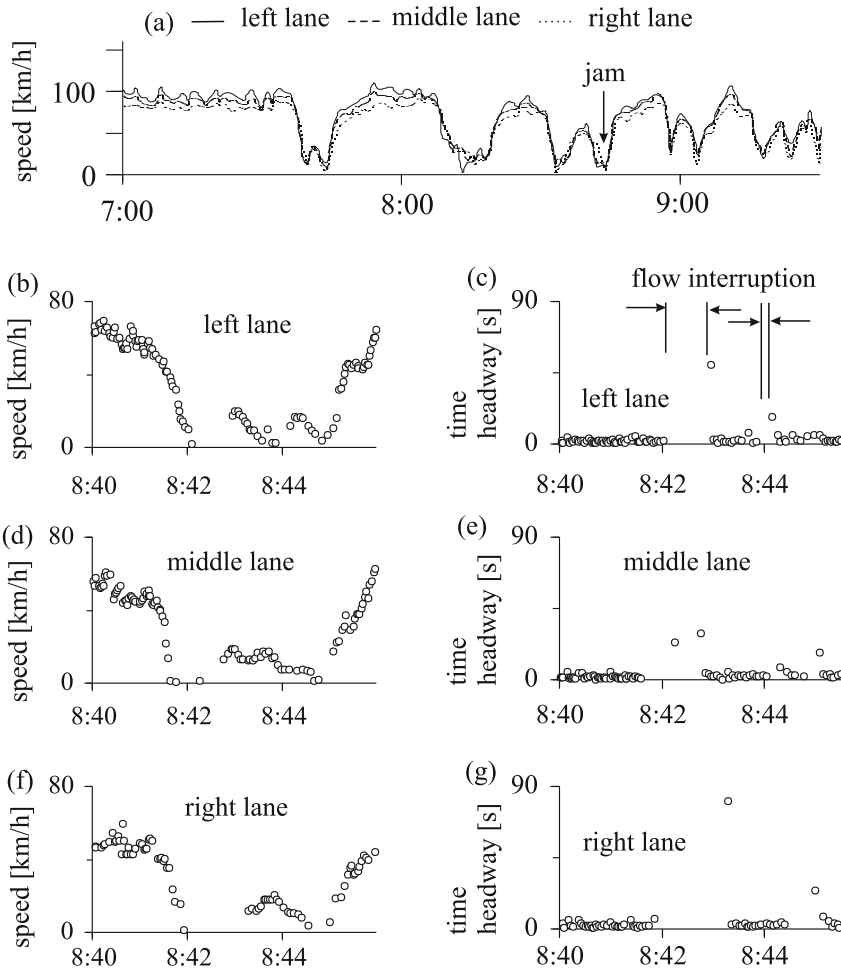
$$I_s = \frac{\tau_{\max}}{\tau_{\text{del}}^{(\text{acc})}(0)} \gg 1, \quad (2)$$

where $\tau_{\max}$ is the maximum time headway between two vehicles within a wide moving jam ($\tau_{\max} \leq \tau_j$).

For this and the following consideration, we should define the term *time headway* between vehicles. At a given time instant $t = t_1$, time

headway (net time gap) $\tau(t_1)$ is *defined* as a time it takes for a vehicle to reach a freeway location at which the bumper of the preceding vehicle is at the time instant t_1 . In single-vehicle data measured at a road location through road detectors (Figs. 10 and 11), t_1 is the time instant at which the preceding vehicle leaves the detector whose location is therefore related to the location of the preceding vehicle in the definition of time headway. Time headway is equal to $\tau(t_1) = t_2 - t_1$, where t_2 is the time instant at which the vehicle front has been recorded at the detector.

Under condition (2), there are at least several vehicles within the jam that are in a standstill, or if they are still moving, it is only with a negligible low speed in comparison with the speed in the jam inflow and outflow. These vehicles separate vehicles accelerating at the downstream jam front from vehicles decelerating at the upstream jam front. In other words, the jam outflow does not depend on the jam inflow. For



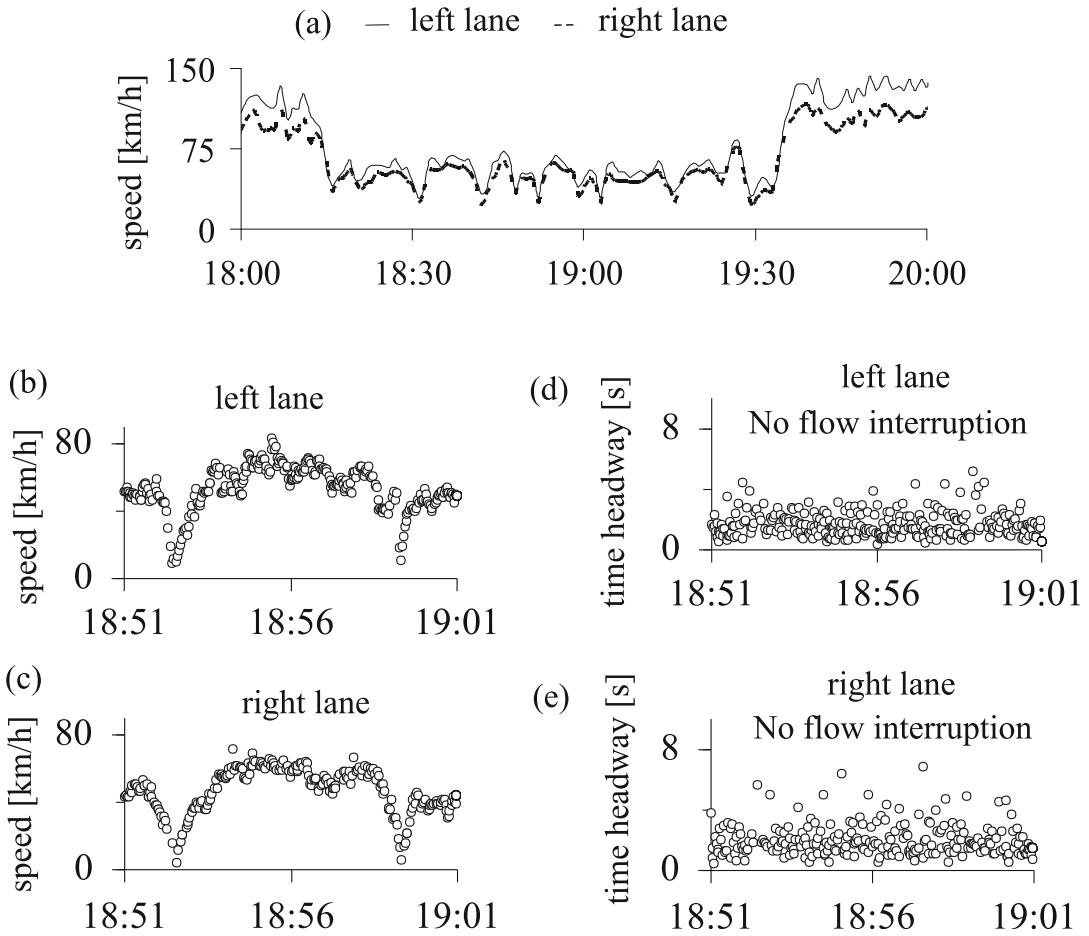
Spatiotemporal Features of Traffic Congestion, Fig. 10 Measured single-vehicle data analysis: (a) overview of measured data (1-min average data). (b–g) Single-vehicle data for speed in different highway lanes for a wide

moving jams (left figures) labeled by “jam” in (a) and the related time headway (right figures). Data has been measured through induction loop detectors installed at a highway location. (Adapted from Kerner et al. 2006b)

this reason, this jam propagates through a bottleneck while maintaining the mean downstream jam front velocity. In accordance with the macroscopic objective spatiotemporal criteria [J] and [S], this jam is a wide moving jam. Thus, we can assume that the traffic flow interruption effect can be used as a criterion to distinguish the synchronized flow and wide moving jam phases in single-vehicle data. This is possible even if data is measured at a single freeway location.

The interruption of traffic flow within a wide moving jam can also be found in *empirical* single-

vehicle data. In an example shown in Fig. 10b–g, flow interruption effect occurs two times during upstream jam propagation through road detectors (these time intervals are labeled “flow interruption” in Fig. 10c). The values $\tau_{\max}$ for the first flow interruption intervals within the wide moving jam are equal to approximately 50 s in the left lane, 24 s in the middle line, and 80 s in the right line. These values $\tau_{\max}$ satisfy criterion (2). This means that traffic flow is discontinuous within the moving jam, i.e., this moving jam can be associated with the wide moving jam phase. Later, some vehicles within the jam have time headway of



Spatiotemporal Features of Traffic Congestion, Fig. 11 Measured single-vehicle data analysis: (a) overview of measured data (1-min average data). (b–e) Single-vehicle data for speed in different highway lanes for

a sequence of narrow moving jams (left figures) and the related time headway (right figures). Data has been measured through induction loop detectors installed at a highway location. (Adapted from Kerner et al. 2006b)

about 4 s or longer associated with moving blanks (Kerner et al. 2006a, b).

In another example shown in Fig. 11b–e, there are two moving jams. However, rather than wide moving jams, these moving jams should be classified as narrow moving jams. A narrow moving jam is a moving jam, which consists of the jam fronts only. Narrow moving jams are associated with the synchronized flow phase. This is because there are no traffic flow interruption intervals which a narrow moving jam. Indeed, in empirical observations upstream and downstream of the jam, as well as within the jam, there are many vehicles that traverse the induction loop detector.

There is no qualitative difference in time headway for different time intervals associated with these narrow moving jams and in traffic flow upstream or downstream of the jams (Fig. 11d, e).

This can be explained if it is assumed that each vehicle that meets a narrow moving jam, which consists of the jam fronts only, can nevertheless accelerate later almost without any time delay within the jam. Within the upstream front, vehicles must decelerate to a very low speed. However, the vehicles can accelerate almost immediately at the downstream jam front. These assumptions are confirmed by single-vehicle data shown in Fig. 11d, e, in which time intervals

between different measurements of time headway for different vehicles exhibit the same behavior outside the jams and within the jams. Even if within a narrow moving jam there are vehicles that are stopped, the condition

$$I_s \sim 1 \quad (3)$$

can be satisfied. Under this condition, there is no flow interruption within this jam. Thus, regardless of these narrow moving jams, traffic flow is not discontinuous, i.e., the narrow moving jams are associated with the synchronized flow phase.

This single-vehicle analysis enables us to assume that congested traffic in Fig. 10b–g is associated with a wide moving jam. In contrast, congested traffic in Fig. 11b–e is associated with the synchronized flow phase:

- We must distinguish between narrow moving jams and wide moving jams (Kerner 2004b, 2009c):
 - (i) A narrow moving jam belongs to the synchronized flow phase of congested traffic.
 - (ii) A wide moving jam belongs to the wide moving jam phase of congested traffic.

We note that a detailed discussion of the physics of the emergence of narrow moving jams will be done in section “Moving Jam Emergence in Framework of Three-Phase Traffic Theory” below.

It should be stressed that for a proof of the microscopic criterion for the traffic phases (2), we should compare results of phase identification based on this criterion in measured single-vehicle data with the traffic phase definitions, i.e., with the macroscopic criteria for the traffic phases [J] and [S] (section “Introduction”). However, such an empirical proof can be possible, if the single-vehicle data associated with a congested pattern in the neighborhood of a bottleneck is measured at many freeway locations including locations upstream and downstream of this bottleneck; this is because only in this case can we prove the behavior of the downstream front of the congested pattern at the bottleneck. Unfortunately, we do not have such measured single-vehicle data. For this reason, we can make a proof of the microscopic

criterion for the traffic phases based on numerical simulations in the framework of the three-phase theory only. This proof, which is based on numerical simulations of the Kerner-Klenov stochastic three-phase traffic flow model (see Appendix A of the book Kerner (2017a) as well as the review article Kerner (2018a) of this Encyclopedia), has been performed in Kerner et al. (2006a) (Fig. 12).

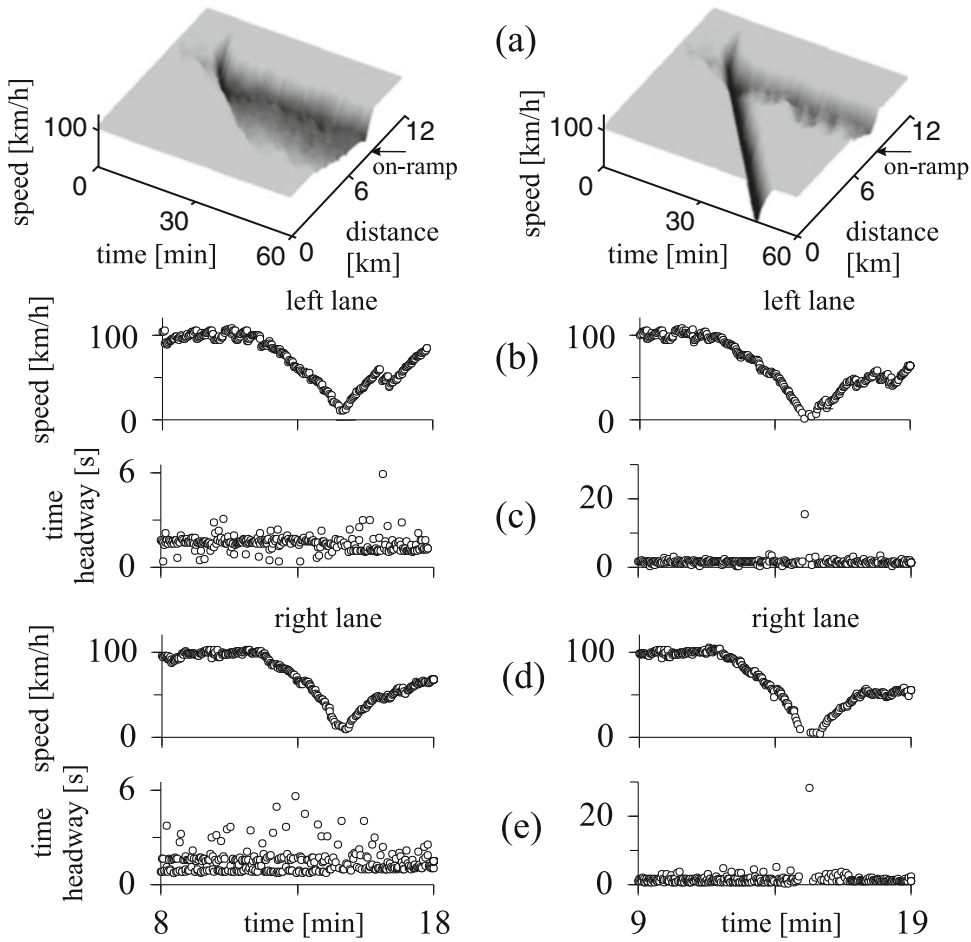
The results of numerical simulations (Fig. 12) allow us to conclude that the microscopic criterion (2) enables us indeed to distinguish the wide moving jam and synchronized flow phases in single-vehicle data measured at one road location.

Moving Blanks Within Wide Moving Jams

The spatiotemporal criterion for a wide moving jam [J] of section “Introduction” that defines the wide moving jam phase in congested traffic can be explained by a traffic flow interruption that occurs when vehicles are in a standstill within the jam. As mentioned, a sufficient criterion for this flow interruption within the jam is given by formula (2) (Kerner et al. 2006b).

The flow interruption effect is a general effect for each wide moving jam (see examples of flow interruption intervals labeled by “flow interruption” in Figs. 10c and 13a). For this reason, criterion (2) can be considered as a microscopic criterion for the wide moving jam phase (more detailed explanations of flow interruption intervals and criterion (2) for wide moving jams appear in Kerner (2009a) and section “Congested Patterns at Heavy Bottlenecks”). Between flow interruption intervals within the wide moving jam shown in Fig. 10b–g, vehicles within the jam exhibit time headway about 1.5–7 s (Fig. 13b). The time headway is considerably shorter than flow interruption intervals (about 20 s and longer) in Fig. 10c, e, g. The time headway is related to low speed states measured at detectors within the jam (Fig. 10b, d, f).

To understand these low speed states, note that when vehicles meet the wide moving jam, firstly the vehicles decelerate to a standstill at the upstream jam front. As a result, the first flow interruption interval in all lanes appears (Figs. 10c, e, g and 13a). Net distances (space gaps) between these vehicles can be very different, and the mean space



Spatiotemporal Features of Traffic Congestion, Fig. 12 Comparison of microscopic criterion with macroscopic spatiotemporal objective criteria [J] and [S] for the phases in congested traffic of identical vehicles (results of model simulations). Left figures are related to a narrow moving jam. Right figures are related to a wide moving jam. (a) The narrow moving jam is caught at an on-ramp bottleneck that is associated with the synchronized flow definition [S] (left); in contrast, the wide moving jam

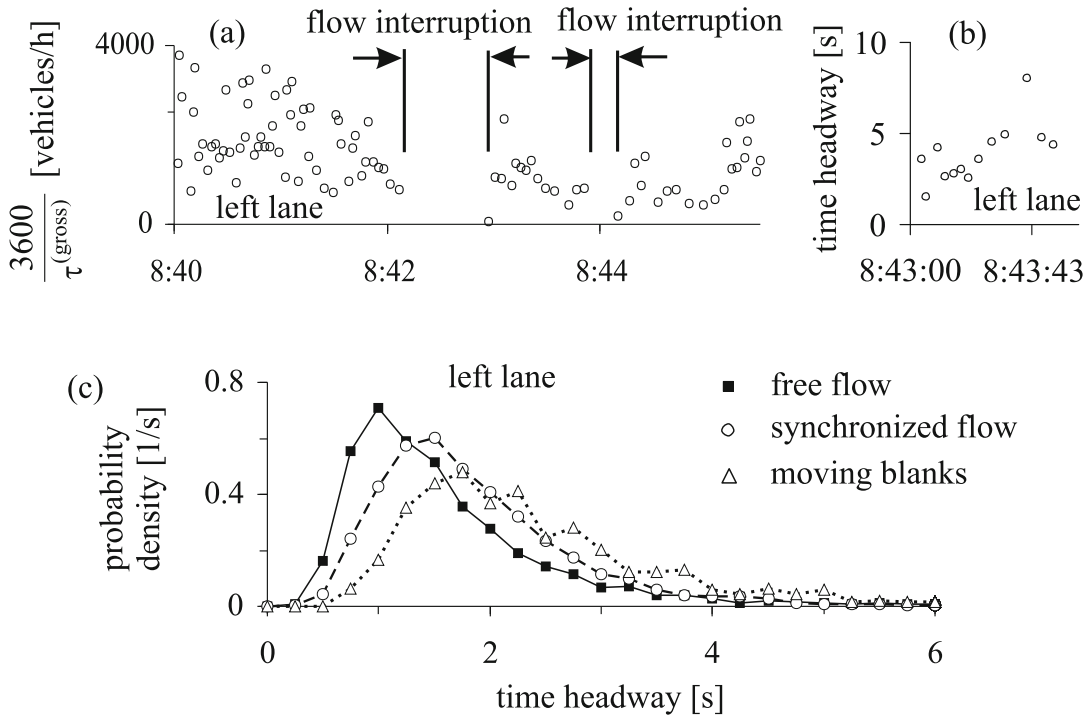
propagates through the bottleneck that is associated with the wide moving jam definition [J] (right). Vehicle speed on the main road in space and time that is averaged across two lanes. (b, d) Single-vehicle data for speed in the left (b) and in right lanes (d) at location 50 m downstream of the end of the on-ramp merging region. (c, e) Time headway associated with (b) and (d), respectively. (Adapted from Kerner et al. 2006a)

gap can exceed a minimum (safe) space gap considerably, i.e., regions with no vehicles can appear within the jam. These regions with no vehicles are called *blanks* within the wide moving jam.

Later, vehicles move covering these blanks. This low-speed vehicle motion is responsible for low speed states mentioned above (Fig. 10b, d, f). Consequently, due to this vehicle motion, new blanks between vehicles occur upstream, i.e., the blanks move upstream within the jam. Then, other

vehicles within the jam that are upstream of these blanks begin also to move covering the latter blanks. This leads to moving blanks that propagate upstream within the jam.

Thus, we define a *moving blank* as a blank between vehicles, which moves upstream due to vehicle motion within a wide moving jam. The vehicle motion occurs at a low vehicle speed, i.e., the vehicle motion creating moving blanks within the jam is related to low speed



Spatiotemporal Features of Traffic Congestion, Fig. 13 Measured microscopic characteristics of traffic phases associated with single-vehicle data shown in Fig. 10b–g: (a) time distributions of the value $3600/\tau^{(gross)}$ in the left lane ($\tau^{(gross)}$ is the gross time gap between vehicles). (b) Time headway as a time-function

associated with moving blanks in the left lane related to the jam shown in Fig. 10b. (c) Probability density of time headway in the left lane for free flow (solid curve), synchronized flow (dashed curve), and moving blanks within wide moving jams (dotted curve). (Adapted from Kerner et al. 2006b)

states discussed above (Fig. 10b, d, f); these low speed states are the reason for moving blank emergence within the jam. This explains why moving blanks are associated with low speed states within the jam.

Time headway (net time gap) $\tau^{(net)}$ between vehicles in these low speed states with moving blanks is about 1.5–7 s (Fig. 13b). Because these low speed states are the reason for moving blanks, we will refer time headway in the low speed states to moving blanks. Thus, the term *time headway for moving blanks* means time headway in the low speed states causing upstream motion of blanks.

To explain a possible short time headway associated with moving blanks observed in empirical data (about 1.5 s), note that when vehicles begin to cover a blank within a wide moving jam, they start to move from a standstill within the jam. Thus, some of the vehicles can move

with time headway between each other that is close to a time delay in vehicle acceleration from a standstill within the jam whose average value is $\tau_{del}^{(acc)}(0) \approx 1.5 - 2$ s.

Time headway in low speed states within the jam, i.e., time headway associated with moving blanks (about 1.5–7 s in Fig. 13b) must be distinguished from long time intervals of traffic flow interruption within a wide moving jam (about 20 s and longer) (Figs. 10c, e, g and 13a). The flow interruption intervals within the jam occur when some of the vehicles within the jam do not move. This vehicle stoppage does not indicate whether there are some blanks between these vehicles or not. This is because during these long flow interruption intervals, space gaps between vehicles that are stopped within the jam might be quite small.

Thus, at a fixed highway location within a wide moving jam, long time intervals, in which no

vehicle passes the location (flow interruption intervals), alternate with time intervals in which vehicles pass the location with low speeds and time headway of about 1.5–7 s associated with moving blanks within the jam. This empirical result could be associated with the following microscopic jam space structure: highway regions in which vehicles are in a standstill (traffic flow interruption) alternate with highway regions in which blanks moving upstream (moving blanks) occur resulting from vehicle motion with low speeds.

Empirical data show that low speed states within wide moving jams associated with moving blanks are not necessarily synchronized between different lanes (Fig. 10b, d, f).

Moving blanks within a wide moving jam resemble electron holes in the valence band of semiconductors. As the moving blanks that propagate upstream appear due to downstream vehicle motion within the jam, so appearance of electron holes moving with the electric field results from electron motion against the electric field in the valence band of semiconductors. However, an electron hole is associated with a single vacancy for one electron in the valence band. In contrast, there can be wide (in the longitudinal direction) moving blanks associated with a “vacancy” for two or more vehicles within a wide moving jam. This moving blank can be split into two or more moving blanks during upstream blanks motion. On the other hand, several adjacent moving blanks can merge into one moving blank (a more detailed consideration of the spatiotemporal dynamics of moving blanks and flow interruption intervals appears in section “[Congested Patterns at Heavy Bottlenecks](#)”).

One of the microscopic characteristics of the traffic phases is time headway distributions. In particular, it has been found that the mean time headway for free flow is shorter than for congested traffic (see references in Helbing 2001). This can be explained at least by the following peculiarities of free flow: (i) A vehicle can come very close to the preceding vehicle just before the vehicle passes it. However, a road detector cannot recognize whether a very short time headway that has been measured is related to car-following or to this vehicle passing effect. (ii) A vehicle can easily change

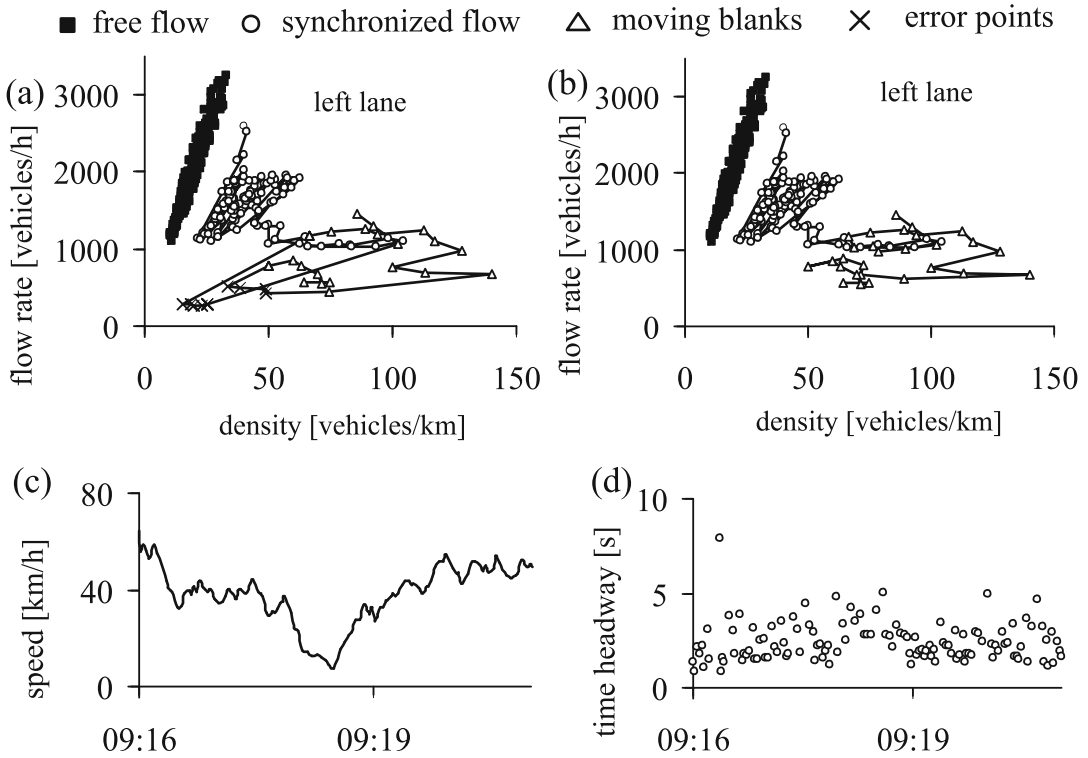
the lane, if the preceding vehicle decelerates unexpectedly; thus, the vehicle can choose a shorter time headway in car-following than in congested traffic. (iii) There is a frustration effect in congested traffic that can be responsible for some long time headway (Helbing 2001).

There can also be another reason for shorter mean time headway in free flow in comparison with mean time headway in congested traffic. It has been found that there are different distributions of time headway in synchronized flow and moving blanks associated with wide moving jams (Fig. 13c). It can be assumed that as long as vehicles are within a wide moving jam, the vehicles have no hurry while covering blanks with the jam.

Traffic Phases in Flow–Density Plane

There are a huge number of publications associated with an analysis of measured traffic data in the flow–density plane (see references in May (1990), Helbing (2001), and Chowdhury et al. (2000)). One of the aims of this analysis is to find criteria for a classification of qualitatively different traffic states and phases associated with congested traffic (see, e.g., Neubert et al. 1999, Knospe et al. 2002, and Knospe et al. 2004). However, as mentioned above, traffic is a complex process in space and time, i.e., rather than a data analysis in the flow–density plane, to find qualitatively different traffic states and phases, one should firstly study spatiotemporal features of traffic congested patterns. Otherwise, invalid conclusions about features of traffic states and phases can be made very easily.

To illustrate this, we consider single-vehicle characteristics of low vehicle speeds within a wide moving jam associated with moving blanks (cross and triangle points) with free flow (black squares) and synchronized flow (circles) in the flow–density plane (Fig. 14a). Note that data used for the calculation of moving blanks shown in Fig. 14a is related to a wide moving jam shown in Fig. 10b–g. In contrast, data used for the calculation of synchronized flow shown in Fig. 14a is related to low speed data shown in Fig. 14c, d, which does not include flow interruption intervals (Fig. 14d).



Spatiotemporal Features of Traffic Congestion, Fig. 14 Empirical characteristics of traffic phases: (a) Traffic phases in the flow–density plane for data in the left lane. (b) The same traffic phases as those in (a), however, with *improved* data for moving blanks. (c, d) Single-vehicle speeds (c) and time headway (d) for synchronized flow states used in (a, b) that are associated with

data shown in Fig. 10a. Microscopic data for moving blanks, for synchronized flow, and for free flow used in (a, b) are associated with single-vehicle data, respectively, for wide moving jam shown in Fig. 10b, c, for synchronized flow shown in (c, d), and for free flow shown in Fig. 10a. In (c, d), moving averaging over five vehicles is used. (Adapted from Kerner et al. 2006b)

As well-known (May 1990; Helbing 2001; Chowdhury et al. 2000), in free flow, the flow rate is on average an increasing density function in the flow–density plane (black squares in Fig. 14a). Due to small density states associated with low vehicle speeds within wide moving jams (cross points in Fig. 14a), the flow rate might also be considered on average an increasing density function in the flow–density plane for wide moving jams. For synchronized flow (Fig. 14a), measured points for synchronized flow exhibit a large scattering in the flow–density plane. These results have been used in Neubert et al. (1999), Knospe et al. (2002, 2004) for the definition of a criterion for the traffic phases in measured data of congested traffic based on the flow–density correlation function.

$$c_{\rho,q}(\tau) = \frac{\langle \rho(t)q(t+\tau) \rangle - \langle \rho(t) \rangle \langle q(t+\tau) \rangle}{\sqrt{\langle \rho^2(t) \rangle - \langle \rho(t) \rangle^2} \sqrt{\langle q^2(t) \rangle - \langle q(t) \rangle^2}}. \quad (4)$$

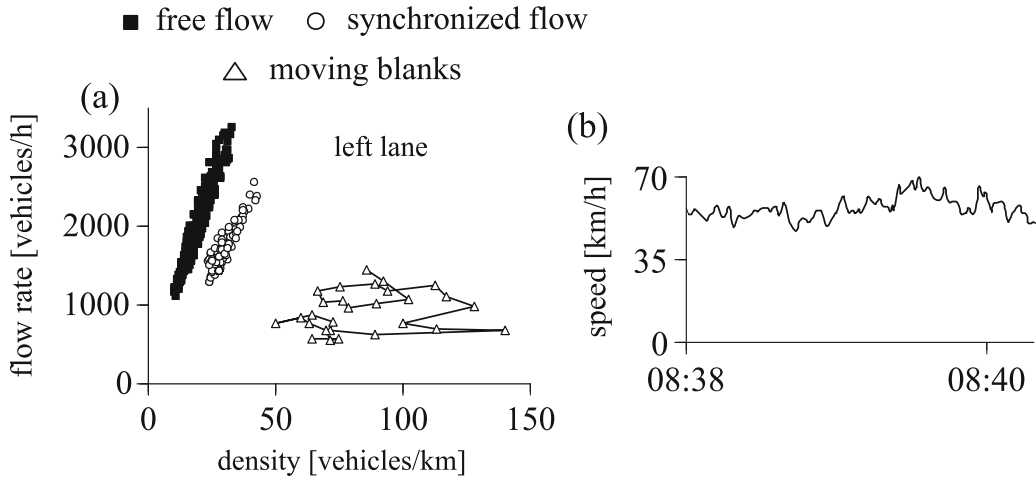
In accordance with Neubert et al. (1999), Knospe et al. (2002, 2004), small flow–density correlations in the data should be associated with synchronized flow, while large flow–density correlations in the data for congested traffic – with wide moving jams. Indeed, for data shown in Fig. 14a, we find that the flow–density correlation coefficient (4) for synchronized flow is a very small value $c_{\rho,q}^{(\text{syn})}(0) \approx 0.06$, whereas for the wide moving jam $c_{\rho,q}^{(\text{jam})}(0) \approx 0.67$, i.e., it is comparable with the correlation coefficient for free flow $c_{\rho,q}^{(\text{free})}(0) \approx 0.97$.

However, these conclusions of the definition for the traffic phases of Neubert et al. (1999), Knospe et al. (2002, 2004) based on the flow–density correlation coefficient (4) are *invalid*. As explained in Kerner et al. (2006b), the increasing density function (Fig. 14a) for wide moving jams in the flow–density plane and the associated large value of flow–density correlations are associated with a *large systematic error* in the data measured at road detectors. This error is due to incorrect density estimation within moving jams made in Neubert et al. (1999), Knospe et al. (2002, 2004). To understand this critical conclusion, recall that there are traffic flow interruption intervals within the wide moving jam (Fig. 10b–g). This leads to an incorrect calculation of the average speed v within the jam. The average speed measured by a detector can be related to a time interval, which includes at least one traffic flow interruption interval. In this case, the measured average speed is considerably higher than the real average speed within the jam. This is because vehicles do not move during traffic flow interruption intervals. As a result, the subsequent density estimation through the formula $\rho = q/v$ (q the flow rate) leads to very small densities within the jam. However, the real density of vehicles stopped within the jam during traffic flow interruption is very large. Thus, the measured data associated with the traffic flow interruption effect within wide moving jams cannot be used for density estimation within the jams. To estimate real vehicle density within a wide moving jam during time intervals that include flow interruption intervals, the density definition (number of vehicles per freeway length at a given time moment), i.e., the *spatial averaging*, should be used, rather than density estimation through the use of the formula $\rho = q/v$ in which the flow rate and average speed are related to the *time averaging* of vehicles passing a detector during a given time interval. The conclusion about the large systematic error in density estimation within wide moving jams is confirmed by a numerical analysis discussed in section “Numerical Simulations of Moving Blanks Within Wide Moving Jams.”

This analysis shows also that within wide moving jams, density estimation $\rho = q/v$ exhibits a relative small error during vehicle motion associated with moving blanks within the jams. This is because in this case the associated time headway is related to vehicles moving within the jams. If only these single-vehicle data is used for density estimation, then points for moving blanks within the jam appear that are associated with random transformations in different directions in the flow–density plane (triangle points in Fig. 14b) rather than an increasing function of the flow rate with density (cross and triangle points in Fig. 14a). This is explained by low vehicle speed states that are associated with non-regular vehicle motion covering blanks within the wide moving jam. If error measurement points are removed (Fig. 14b), then the flow–density correlation coefficient for the wide moving jam $c_{\rho,q}^{(\text{jam})}(0) \approx 0.18$, i.e., it is small. This means that in contrast with Neubert et al. (1999), Knospe et al. (2002, 2004), no valid traffic phase definition based on the flow–density correlations in congested traffic can be made.

In addition, we should mention that synchronized flow states can overlap states associated with moving blanks within a wide moving jam as this is shown in Fig. 14b. This is because the speed in synchronized flow states can be as low as the speed in states associated with moving blanks within the jam. Thus, in a general case, low speed states associated with moving blanks within wide moving jams cannot also be used for clearly distinguishing the synchronized flow and wide moving jam phases in congested traffic.

Moreover, the result mentioned above that synchronized flow exhibits a small flow–density correlation, as this is the case for measured data shown in Fig. 14b, c, is *not* a general one. An empirical example of synchronized flow, for which the flow–density correlation coefficient $c_{\rho,q}^{(\text{syn})}(0) \approx 0.91$ is as large as the one for free flow, is shown in Fig. 15. Thus, we see that in contrast with Neubert et al. (1999), Knospe et al. (2002, 2004), a large value of the flow–density correlation cannot be used as the criterion for distinguishing free flow from synchronized flow.



Spatiotemporal Features of Traffic Congestion, Fig. 15 Empirical characteristics of traffic phases: (a) Traffic phases in the flow–density plane for data in the left lane with improved data for moving blanks. Data for

free flow and wide moving jam are the same as those in Fig. 14b. (b) Single-vehicle speeds for synchronized flow used in (a) are associated with data shown in Fig. 10a. In (a), moving averaging over five vehicles is used

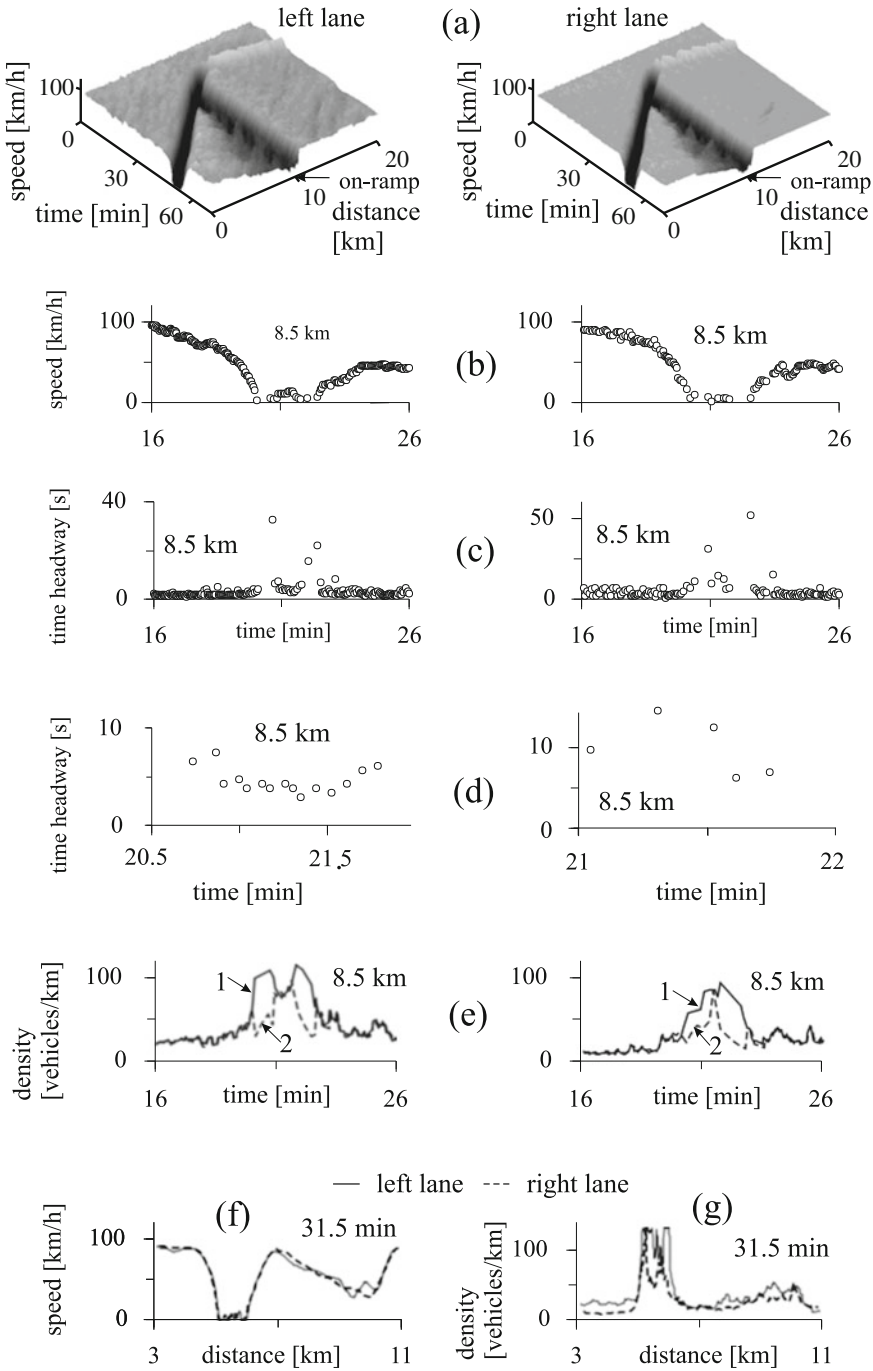
Numerical Simulations of Moving Blanks Within Wide Moving Jams

To understand features of propagation and evolution of moving blanks within a wide moving jam, we consider a moving jam that is induced downstream of an on-ramp bottleneck by applying a short-time and high-amplitude local speed disturbance in heterogeneous free flow consisting of fast and long vehicles (Fig. 16). There are two time intervals in which flow interruption within the jam occurs. The maximum time headway (Fig. 16c) associated with these flow interruption effects satisfy microscopic criterion (2) for wide moving jams (Kerner et al. 2006a, b, 2007; Kerner 2009a) (time delays used in (2) for fast and long vehicles are $\tau_{\text{del}}^{(\text{acc})}(0) \approx 1.77$ and 3.33 s, respectively). As a result, in accordance with the macroscopic spatiotemporal criterion [J] for the wide moving jam phase, this jam propagates through the bottleneck while maintaining the mean velocity of the downstream jam front.

Within the wide moving jam, between time intervals of flow interruptions, there are intervals in which low speed states appear associated with moving blanks (Fig. 16b–d). Simulated time headway for moving blanks in the left lane that changes over time within a range 2.5–7.5 s (Fig. 16d, left) correspond to empirical observed

values (Fig. 13). As can be seen from vehicle trajectories within the jam, these low speed states associated with moving blanks are related to non-homogeneous vehicle motions that cover blanks between vehicles within the jam (Fig. 17). These moving blanks propagate upstream with a negative velocity that is on average equal to the characteristic speed of the downstream jam front. Spatial and temporal speed distributions in low speed states associated with moving blanks exhibit complex variations within the jam, which are different in the left and right lanes (Fig. 17b, c); these variations seem to correspond to a random vehicle speed behavior within the jam.

In accordance with empirical results of section “Traffic Phases in Flow–Density Plane,” due to flow interruption within the wide moving jam shown in Fig. 16a–c, there is a large error in density distributions calculated through the formula $\rho = q/v$ at large vehicle density (curves 2 in Fig. 16e) in comparison with density distributions that are found based on the density definition (vehicles per freeway length) (curves 1). Indeed, curves 2 in Fig. 16e show a significant density underestimation within wide moving jams. These error points (crosses in Fig. 18a) explain the systematic error in the empirical studies of states within wide moving jams made in Neubert



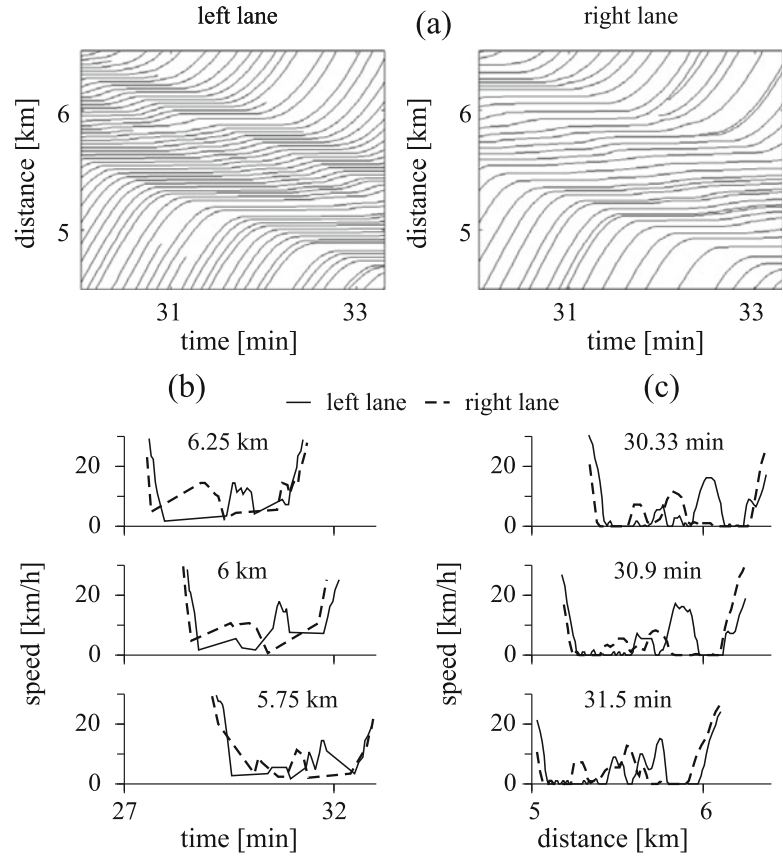
Spatiotemporal Features of Traffic Congestion,

Fig. 16 Simulated characteristics of moving blanks within a wide moving jam propagating through an on-ramp bottleneck. (a) Speed in space and time on the main road. (b–e) Single-vehicle data for time dependencies of speed (b), time headway (c, d), and the associated density distributions (e) calculated through the density

definition (solid curves 1) and, as in empirical Fig. 14a, through the use of formula $\rho = q/v$ (dashed curves 2) at location $x = 8.5$ km. Left and right figures (a–e) are related to the left and right lanes, respectively. (f, g): speed (f) and density (g) related to (a) as functions of distance at time 31.5 min in the left and right lanes. (Adapted from Kerner et al. 2006b)

Spatiotemporal Features of Traffic Congestion,

Fig. 17 Continuation of Fig. 16. Simulated microscopic characteristics of moving blanks within the wide moving jam shown in Fig. 16a: (a) Vehicle trajectories for each fourth vehicle in the left (left) and right (right) lanes. (b, c) Low speed states associated with moving blanks at three different locations (b) and three different time moments (c) in the left and right lanes. (Adapted from Kerner et al. 2006b)

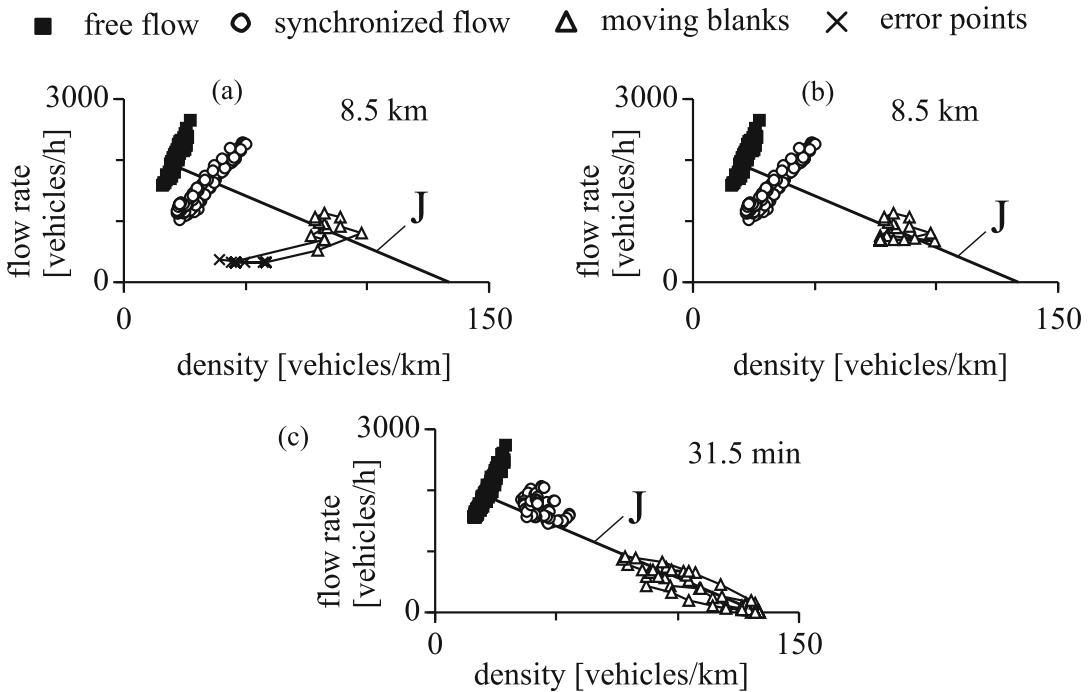


et al. (1999), Knospe et al. (2002, 2004) that has been illustrated in Fig. 14a (compare error points in Figs. 14a and 18a).

Within the regions of moving blanks, the difference between density calculated through the formula $\rho = q/v$ (curves 2 in Fig. 16e) and density found based on the density definition (curves 1) decreases considerably. For this reason, if error points associated with flow interruption within the jam are removed, then remaining points in the flow–density (Fig. 18b) exhibit a qualitative correspondence with the states within the jam calculated through the density definition (Fig. 18c). The latter states are related to the speed and density found at a fixed time moment 31.5 min (Fig. 16f, g). Nevertheless, there are some quantitative differences between speed and density distributions found through detector measurements (Fig. 18b) and through the use of the density definition in Fig. 18c. In particular, there

are points in the flow–density plane found in the density calculation through the density definition (Fig. 18c), which are related to larger density up to the maximum jam density. These points cannot usually be found based on traffic measurements at a detector location (Fig. 18b).

To understand possible scenarios of the emergence of moving blanks within a wide moving jam, spontaneous jam emergence in synchronized flow of a general pattern (GP) at an on-ramp bottleneck has been simulated in heterogeneous flow consisting of passenger cars and long vehicles (Fig. 19). Firstly, an F→S transition occurs spontaneously at the bottleneck. Synchronized flow that appears due to this F→S transition propagates upstream. Moving jams emerge spontaneously in that synchronized flow propagates upstream (in more details, GPs will be considered in section “General Congested Patterns”).



Spatiotemporal Features of Traffic Congestion, Fig. 18 Continuation of Fig. 16. Simulated characteristics of moving blanks within the wide moving jam in the left lane shown in Fig. 16a together with states of free flow and synchronized flow in the flow–density plane. (a, b) States within the jam that are determined at detector at $x = 8.5$ km through the formula $\rho = q/v$ in two cases in

which all states within the jam measured at the detector are shown (a) and error states are removed (b). (c) States within the jam that are determined through the density definition (vehicles per freeway length) at time 31.5 min. J is the line J for the downstream jam front. (Adapted from Kerner et al. 2006b)

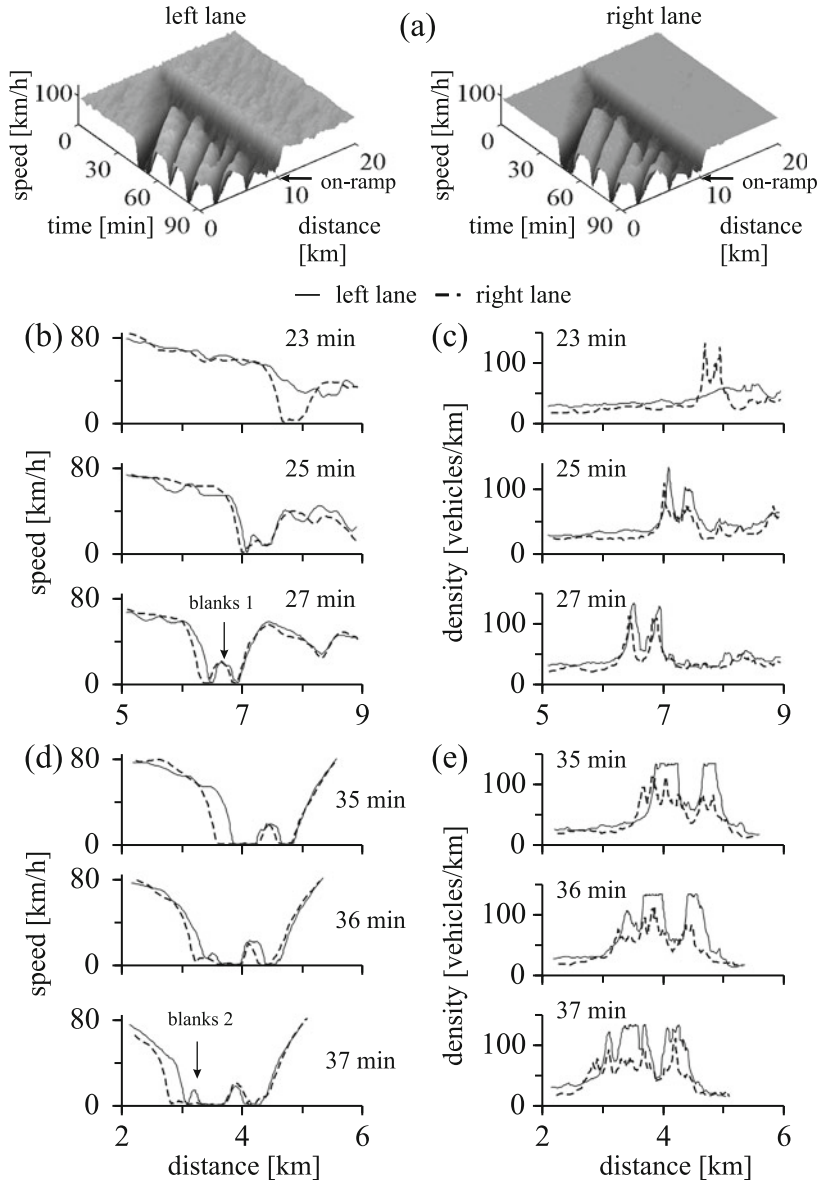
Simulations show that there are the following mechanisms for emergence of moving blanks within wide moving jams: (i) Vehicle lane changing at the upstream jam front. (ii) Vehicles come to a stop at the upstream front of a wide moving jam at different space gaps to the related preceding vehicles. (iii) When vehicles move with low speeds covering downstream blanks within the jam, vehicle lane changing can result in new moving blanks.

In heterogeneous flow, moving blanks emerge most frequently due to vehicle lane changing (item (i) and (iii)). Specifically, a wide moving jam emerges firstly in the right lane only ($t = 23$ min in Figs. 19b, c and 20a). Then, some vehicles change from the right lane to the left lane (arrow 1 in Fig. 20a). On the one hand, this lane changing decreases speed in the left lane. This leads to narrow moving jam formation

in this lane. On the other hand, blanks between vehicles in the right lane increase due to lane changing. Vehicles begin to cover these blanks ($t \approx 24$ min in Fig. 20a). As a result, low speed states within the wide moving jam appear associated with these moving blanks. Later, subsequent lane changing of vehicles from the right to the left lane results in an abrupt decrease in speed in the left lane upstream of the narrow moving jam ($t \approx 25$ min in Fig. 19b, c and arrow 2 in Fig. 20a). Then, this complex spatiotemporal speed distribution in the left lane transforms into a wide moving jam with moving blanks within the jam ($t \approx 27$ min in Fig. 19b, c). Thus, in this scenario, vehicle lane changing from the right to the left lane at the upstream jam front is the main reason for emergence of moving blanks within the wide moving jam (moving blanks labeled “blanks 1” in Fig. 19b).

Spatiotemporal Features of Traffic Congestion,

Fig. 19 Simulated characteristics of moving blanks emergence: (a) Speed in space and time within a general pattern (GP) at an on-ramp bottleneck in the left (left) and right lanes (right). (b–e) Speed (b, d) and density (c, e) at different time moments associated with the GP in (a) in the left and right lanes. In (c, e), for density calculation, the density definition (vehicles per freeway length) is used. (Adapted from Kerner et al. 2006b)



During subsequent propagation of the wide moving jam, new moving blanks emerge (Fig. 19d, e). Firstly, the upstream jam front in the right lane is upstream of the upstream jam front in the left lane ($t = 35$ min, Fig. 19d, e). Due to lane changing of vehicles from the right to the left lane, the upstream jam front locations are synchronized with each other (arrow 3 in Fig. 20b). This lane changing causes emergence of moving blanks within the jam ($t = 37$ min, moving blanks labeled “blanks 2” in Fig. 19d).

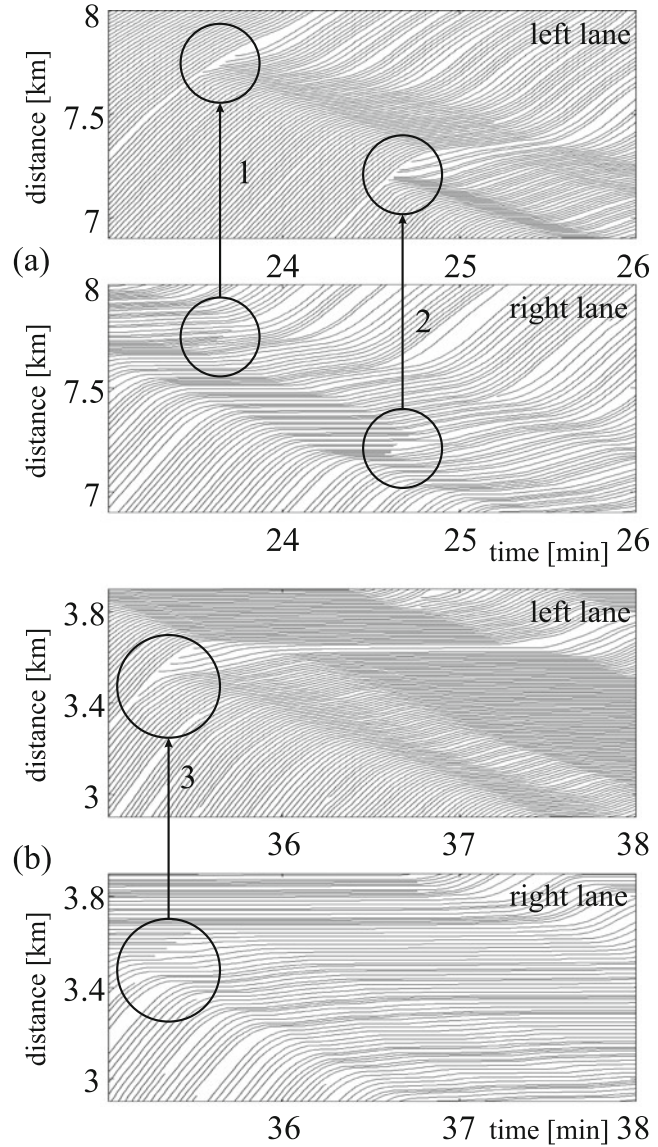
Moving Jam Emergence in Framework of Three-Phase Traffic Theory

The emergence of a wide moving jam in synchronized flow is called an $S \rightarrow J$ transition. In empirical observations, spatiotemporal features of the emergence of the wide moving jam in synchronized flow are as follows (Kerner 1998c, 2004b):

- Firstly, the pinch effect in synchronized flow is realized. The pinch effect is a compression

Spatiotemporal Features of Traffic Congestion,

Fig. 20 Continuation of Fig. 19. Simulated characteristics of emergence of moving blanks in the GP shown in Fig. 19a: (a, b) vehicle trajectories for different time intervals in the left and right lanes. (Adapted from Kerner et al. 2006b)



of synchronized flow with the subsequent emergence of narrow moving jams. The narrow moving jams propagate upstream in synchronized flow. A narrow moving jam results from an S→J instability occurring in the pinch region of synchronized flow. The S→J instability is a growing wave of a local speed reduction in synchronized flow.

- If the narrow moving jam grows continuously over time, the growing narrow moving jam transforms into a wide moving jam. This

means that in this case the S→J instability causes the S→J transition.

- The road location at which the growing narrow moving jam transforms into the wide moving jam determines the upstream boundary of the pinch region of synchronized flow.

Pinch Effect: Narrow Moving Jam Emergence in Synchronized Flow

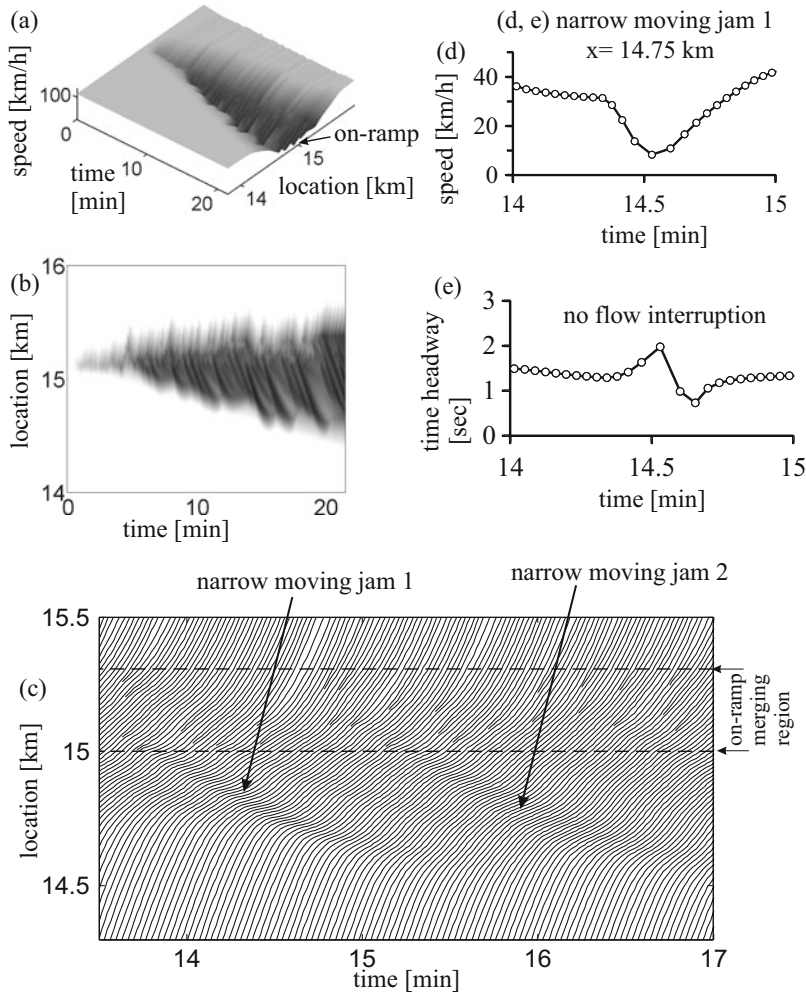
However, the S→J instability should not necessarily lead to the S→J transition. Indeed, in some

cases, *none* of the narrow moving jams occurring due to the pinch effect transforms into a wide moving jam. This case is presented in Fig. 21. Thus, we should distinguish two different consequences of the pinch effect in synchronized flow:

- (i) None of the narrow moving jams occurring due to the pinch effect transforms into a wide

moving jam. This means that the $S \rightarrow J$ instability does not lead to the $S \rightarrow J$ transition (Fig. 21).

- (ii) At least one of the narrow moving jams occurring due to the pinch effect does transform into a wide moving jam. This means that the $S \rightarrow J$ instability does lead to the $S \rightarrow J$ transition. This case will be considered



Spatiotemporal Features of Traffic Congestion, Fig. 21 Simulations of the pinch effect with narrow moving jam emergence, however, *without* subsequent $S \rightarrow J$ transitions. (a, b) Speed in space and time (a) and the same data presented by regions with variable shades of gray (shades of gray vary from white to black when the speed decreases from 90 km/h (white) to zero (black)) (b). (c) Fragments of vehicle trajectories. (d, e) Microscopic vehicle speeds (d) and time headway (e) as time-functions

measured through a virtual detector at road location $x = 14.75$ km in (c). Simulations of the Kerner-Klenov deterministic ATD model (Kerner and Klenov 2006) on a single-lane road with an on-ramp bottleneck at model parameters given in Tables 1 and 2 of Kerner and Klenov (2006). $q_{in} = 1777$ vehicles/h, $q_{on} = 480$ vehicles/h. The beginning and the end of the on-ramp merging region are, respectively, $x_{on} = 15$ km and $x_{on}^{(e)} = 15.3$ km

in sections “[Emergence of Nucleus for S→J Instability and S→J Transition as Spatiotemporal Phenomena](#)” and “[General Congested Patterns](#)” below.

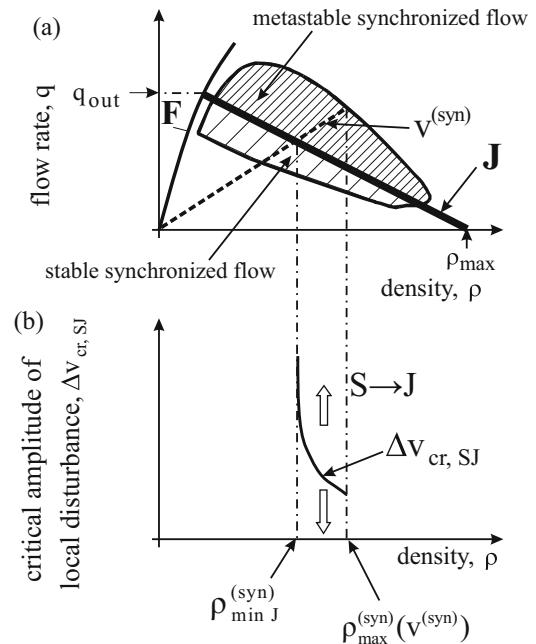
Simulations of the pinch effect in synchronized flow that occurs at an on-ramp bottleneck on a single-lane road (Fig. 21) have been made with the Kerner-Klenov deterministic microscopic ATD model in the framework of the three-phase theory. We can see that narrow moving jams (labeled by “narrow moving jam 1” and “narrow moving jam 2” in Fig. 21c) that emerge in synchronized flow do *not* transform into wide moving jams. To prove that these moving jams are narrow moving jams, in Fig. 21d, e, we show the microscopic speeds and time headway between vehicles moving through the narrow moving jam labeled by “narrow moving jam 1” in Fig. 21c. We can see that the microscopic criterion (2) for a wide moving jam is not satisfied: there is no flow interruption within the narrow moving jam (labeled by “no flow interruption” in Fig. 21e). This means that narrow moving jam 1 belongs to the synchronized flow phase. The analysis made shows that the same conclusion is valid for “narrow moving jam 2” in Fig. 21c. Therefore, we can make the following conclusions:

- The emergence of a narrow moving jam due to the S→J instability should not necessarily lead to the emergence of a wide moving jam (S→J transition).
- The S→J instability should be strictly distinguished from the S→J transition.

Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow
Empirical S→J instability and S→J transition are explained in the three-phase theory through the following hypotheses (Kerner 1998a, c, 1999a):

- (a) All infinite steady states of traffic flow in the flow–density plane that lie on the line J are threshold states for the existence of wide moving jams as well as for wide moving jam emergence (S→J transition) (Fig. 22a).

- (b) The line J intersects the 2D region of steady states of synchronized flow in the flow–density plane: there are synchronized flow steady states below and above the line J (Fig. 22a).
- (c) The line J separates all steady states of synchronized flow in the flow–density plane into two different classes:
- All states of synchronized flow below the line J are stable with respect to moving jam emergence, i.e., no S→J instability and, therefore, no S→J transition are possible within these states (labeled by “stable synchronized flow” in Fig. 22a).



Spatiotemporal Features of Traffic Congestion, Fig. 22 Illustration of hypotheses of three-phase traffic theory about S→J instability and S→J transition: (a) Qualitative presentation of free flow (F), 2D steady states of synchronized flow (2D dashed region), and the line J (line J) in the flow–density plane. (b) Amplitude of critical nucleus $\Delta v_{cr, SJ}$ required for S→J instability as the dependence of vehicle density ρ in different metastable states of synchronized flow of the same vehicle speed $v^{(syn)}$ (dashed line labeled by $v^{(syn)}$ in (a)); arrow labeled by “S→J” symbolizes the growth of the amplitude of a disturbance of a local speed reduction, i.e., the decrease of the minimum speed within the disturbance that becomes a nucleus for the S→J instability. (Adapted from Kerner 1998a, c, 1999a)

- All states on and above the line J are metastable states with respect to wide moving jam existence and emergence ($S \rightarrow J$ transition) (labeled by “metastable synchronized flow” in Fig. 22a).
- (d) There is a critical amplitude of a disturbance of a local speed reduction in metastable synchronized flow leading to an $S \rightarrow J$ instability. The critical amplitude of the speed disturbance denoted by $\Delta v_{cr,SJ}$ can be considered a critical nucleus for the $S \rightarrow J$ instability (Fig. 22b).
 - (e) At a given synchronized flow speed, the larger the density in synchronized flow states that are above the line J in the flow–density plane, the smaller the critical nucleus $\Delta v_{cr,SJ}$ leading to the $S \rightarrow J$ instability (Fig. 22b) and, therefore, the larger the probability of the $S \rightarrow J$ instability.
 - (f) At a given flow rate in synchronized flow, the smaller the speed in metastable synchronized flow, the smaller the critical nucleus $\Delta v_{cr,SJ}$ in synchronized flow leading to the $S \rightarrow J$ instability and, therefore, the larger the probability of the $S \rightarrow J$ instability.
- (i) A driver decelerates with a delay caused by the driver reaction time. Nevertheless, due to the speed adaptation effect within 2D states of synchronized flow, the vehicle is able to adapt the speed to the speed of the preceding vehicle. Then, rather than an $S \rightarrow J$ instability, a new synchronized flow state with a lower speed can be formed. In other words, the $S \rightarrow J$ instability does not occur when within the local speed disturbance the speed adaptation effect within 2D states of synchronized flow is on average stronger than the over-deceleration effect.
 - (ii) Due to the possibility of the speed adaptation effect within 2D states of synchronized flow, a driver decelerates trying to adapt its speed of the preceding vehicle. Nevertheless, due to the driver reaction time, the vehicle decelerates stronger than it would be needed for speed adaptation. As a result, the vehicle speed becomes lower than the speed of the preceding vehicle. If each of the following vehicles decelerates also to a lower speed than the speed of the associated preceding vehicle, the $S \rightarrow J$ instability is realized. In other words, the $S \rightarrow J$ instability does occur when within the local speed disturbance the speed adaptation effect within 2D states of synchronized flow is on average weaker than the over-deceleration effect.

In the three-phase theory (Kerner 1998a, c, 1999a, 2004b), the competition between the driver speed adaptation effect within 2D states of synchronized flow (driver speed adaptation effect has been in details considered in the Encyclopedia article Kerner 2018a) and the driver over-deceleration effect of Herman, Gazis, Montroll, Potts, Rothery, and Chandler (Gazis et al. 1961, 1959; Chandler et al. 1958; Herman et al. 1959) is responsible for the $S \rightarrow J$ instability. To explain this competition between speed adaptation and over-deceleration, we assume that a vehicle moving in a metastable synchronized flow state (a state that is above the line J in Fig. 22a) decelerates unexpectedly. If the following vehicle cannot pass this preceding vehicle, the vehicle should decelerate. As a result, a disturbance of a local speed reduction occurs in synchronized flow. The speed within the disturbance is lower than the initial synchronized flow speed. Because there is a driver reaction time, each vehicle that reaches the local disturbance decelerates with a time delay. There are two possibilities in this case (Kerner 2004b):

Thus, if there is a disturbance with a local speed reduction in synchronized flow, there can be two opposite effects. Due to the speed adaptation effect, there is a tendency toward another state within 2D states of synchronized flow. In contrast, due to over-deceleration, there is also a tendency toward the emergence of a moving jam.

As mentioned, synchronized flow associated with a point in the flow–density plane that lies on or above the line J (Fig. 22a) is in a *metastable state* with respect to moving jam emergence. Due to this synchronized flow metastability, in the three-phase theory, it is required that to initiate an $S \rightarrow J$ instability, a nucleus for the $S \rightarrow J$ instability should appear in a metastable state of synchronized flow (“metastable synchronized flow” in Fig. 22a) (Kerner 1998a, c, 1999a). To understand the term

nucleus for the S→J instability, we consider a disturbance with a local speed reduction occurring in metastable synchronized flow of a synchronized flow speed denoted by $v^{(\text{syn})}$ (dashed line in Fig. 22a shows different points in the flow–density plane at the same speed $v^{(\text{syn})}$). We denote the average speed within this disturbance by $v_{\text{dis}}^{(\text{syn})}$.

When the amplitude of the disturbance $\Delta v_{\text{SJ}} = v^{(\text{syn})} - v_{\text{dis}}^{(\text{syn})}$ is equal to or exceeds a critical amplitude denoted by $\Delta v_{\text{cr,SJ}}$ in Fig. 22b, then and only then the S→J instability is realized (symbolically, this S→J instability is shown by up-arrow “S→J” in Fig. 22b). In this case, the speed adaptation effect within 2D states of synchronized flow is on average weaker than the over-deceleration effect. Thus, a disturbance of the local speed reduction in synchronized flow that amplitude satisfies condition

$$\Delta v_{\text{SJ}} \geq \Delta v_{\text{cr,SJ}} \quad (5)$$

is a nucleus for the S→J instability: the nucleus initiates the development of the S→J instability. Due to the S→J instability, while propagating upstream in the synchronized flow, the nucleus grows in the amplitude: the vehicle speed decreases within such a growing nucleus. The growth of the nucleus required for the S→J instability results in the formation of a narrow moving jam (Kerner 2004b).

Contrarily, if $\Delta v_{\text{SJ}} < \Delta v_{\text{cr,SJ}}$, then the disturbance decays over time (symbolically, the decay of the disturbance is shown by down-arrow in Fig. 22b). In this case, the speed adaptation effect within 2D states of synchronized flow is on average stronger than the over-deceleration effect. At a given synchronized flow speed, the larger the density of synchronized flow, the smaller should be the amplitude of the critical nucleus $\Delta v_{\text{cr,SJ}}$ for the development of the S→J instability (Fig. 22b). This is the hypothesis (e) mentioned above.

Thus, the three-phase theory makes the following statement for the S→J instability in synchronized flow:

- The S→J instability is realized in a metastable synchronized flow with respect to the S→J

transition under condition that a disturbance of the local speed reduction in the synchronized flow appears that is a nucleus required for the S→J instability.

There can be various sources for a nucleus that occurrence and subsequent growth in metastable synchronized flow lead to an S→J instability (Kerner 2004b): unexpected braking of a vehicle in synchronized flow, lane changing within synchronized flow on the main road, fluctuations in flow rates, vehicle merging from other roads (e.g., at bottlenecks), etc.

We have mentioned that due to an S→J instability, a narrow moving jam emerges and propagates in synchronized flow. An S→J transition that can result from the S→J instability is the transformation of the growing narrow moving jam into a wide moving jam. However, as emphasized in section “Pinch Effect: Narrow Moving Jam Emergence in Synchronized Flow,” the S→J instability that leads to the emergence of narrow moving jams in synchronized flow should not necessarily lead to the wide moving jam emergence (S→J transition). This result (see Fig. 21) together with the microscopic analysis of narrow and wide moving jams made in section “Microscopic Criterion for Traffic Phases in Congested Traffic” allows us to make the following conclusions:

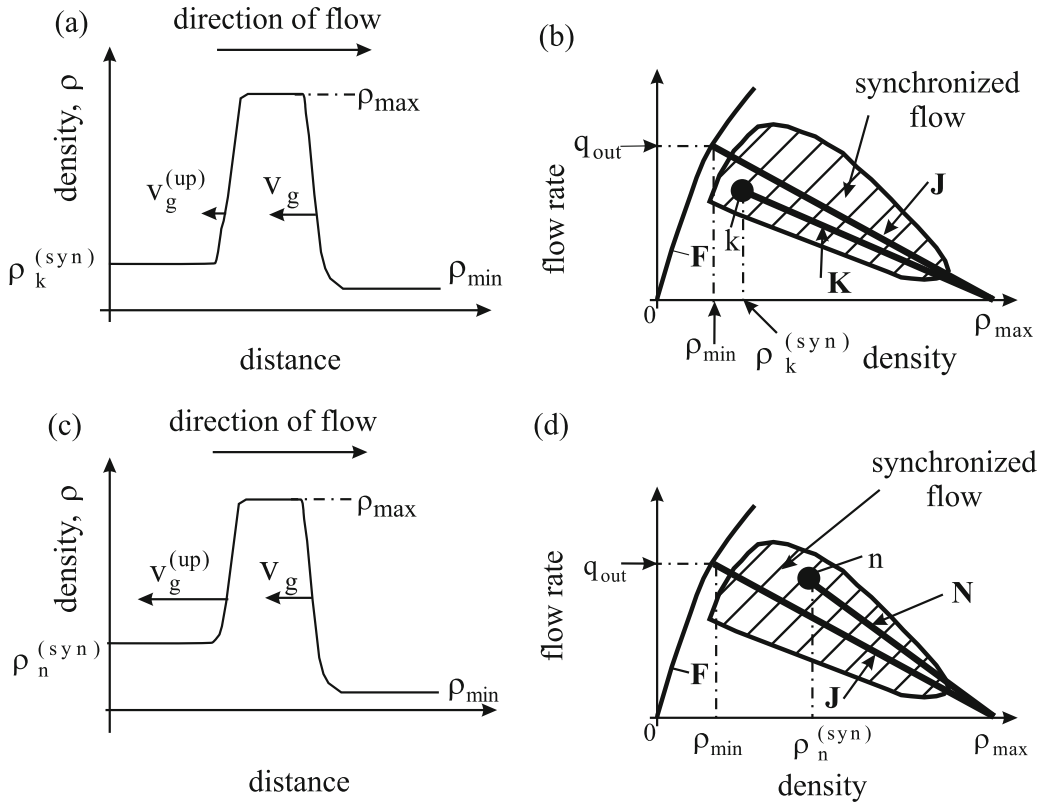
- A narrow moving jam belongs to the synchronized flow phase.
- Contrarily, a wide moving jam belongs to the wide moving jam phase.
- Therefore, the narrow moving jam should be strictly distinguished from the wide moving jam.
- Respectively, the S→J instability (emergence of a narrow moving jam) should be strictly distinguished from the S→J transition (emergence of a wide moving jam).

We continue a consideration of this subject in next section “Emergence of Nucleus for S→J Instability and S→J Transition as Spatiotemporal Phenomena.”

For a more detailed explanation of the metastability of synchronized flow with respect to an S→J transition, let us assume that a state of synchronized flow directly upstream of a wide moving jam shown in Fig. 23a is associated with a point k in the flow–density plane (Fig. 23b). This point is below the line J (Fig. 23b). Because the velocity of the upstream front of the wide moving jam $v_g^{(up)}$ equals the slope of the line K (from a point k in synchronized flow to the point $(\rho_{max}, 0)$), the absolute value $|v_g^{(up)}|$ is always less than the absolute value $|v_g|$ of the velocity of the downstream front that determines the slope of the line J in the flow–density plane; in other words, the formula $|v_g^{(up)}| < |v_g|$ is valid (Fig. 23b). Therefore, the width (in the

longitudinal direction) of the wide moving jam gradually decreases, and the jam dissolves. This means that no wide moving jams can persist continuously. Therefore, all steady states of synchronized flow below the line J are stable with respect to moving jam emergence.

In contrast, we assume now that a state of synchronized flow upstream of another wide moving jam shown in Fig. 23c is associated with a point n in the flow–density plane (Fig. 23d). This state is above the line J (Fig. 23d). In this case, the velocity of the upstream front of the wide moving jam $v_g^{(up)}$ equals the slope of the line N (from a point n in synchronized flow to the point $(\rho_{max}, 0)$), i.e., the absolute value $|v_g^{(up)}|$ is always larger



Spatiotemporal Features of Traffic Congestion, Fig. 23 Continuation of Fig. 22. Explanations of moving jam existence in synchronized flow in three-phase traffic theory: (a, c) Qualitative forms of wide moving jams at two different densities in synchronized flow upstream of the jams, $\rho^{(syn)} = \rho_k^{(syn)}$ (a) and $\rho^{(syn)} = \rho_n^{(syn)}$ (c). (b, d)

Representation of the line J and upstream wide moving jam front (lines K (b) and N (d)) in the flow–density plane for the wide moving jams in (a) and (c), respectively. In (b, d), states for free flow (curve F), line J , and steady states of synchronized flow (2D dashed region) are the same as those in Fig. 22a. (Adapted from Kerner 1998a, c, 1999a)

than the absolute value $|v_g|$ of the velocity of the downstream front; in other words, the formula $|v_g^{(up)}| > |v_g|$ is valid (Fig. 23d). Therefore, the width of the wide moving jam in Fig. 23c should gradually increase. For these reasons, wide moving jams can be formed in states of synchronized flow that lie on or above the line J : these synchronized flow states are metastable with respect to $S \rightarrow J$ transitions. This analysis explains the hypothesis (c) mentioned above.

In synchronized flow states that are on the line J , the velocities of the downstream and upstream fronts of a wide moving jam are equal to each other; this explains the hypothesis (a) that these states are threshold ones for the $S \rightarrow J$ transition (see for more details Sect. 6.3.1 of the book Kerner 2004b).

Emergence of Nucleus for $S \rightarrow J$ Instability and $S \rightarrow J$ Transition as Spatiotemporal Phenomena

As followed from the hypotheses about nucleation nature of moving jam emergence in synchronized flow (Fig. 22) (Kerner 1998a, c, 1999a), an $S \rightarrow J$ instability is realized in a metastable synchronized flow with respect to an $S \rightarrow J$ transition under condition that in the synchronized flow a disturbance of a local speed reduction appears that is a nucleus required for the $S \rightarrow J$ instability.

It should be noted that the hypotheses about moving jam emergence discussed in section “Emergence of Nucleus for $S \rightarrow J$ Instability and $S \rightarrow J$ Transition as Spatiotemporal Phenomena” give only a rough description of the nucleus emergence in synchronized flow. This is because the hypotheses formulate features of hypothetical steady states of synchronized flow. Contrarily, synchronized flow that occurs at a highway bottleneck is essentially nonhomogeneous in space and time. In the nonhomogeneous in space and time synchronized flow, both the emergence of the nucleus and the subsequent development of the $S \rightarrow J$ instability are complex *spatiotemporal* phenomena. Nevertheless, these spatiotemporal phenomena depend considerably on the initial amplitude of the disturbance in synchronized

flow that initiates the phenomena. Here, we illustrate this statement based on numerical simulations of the Kerner-Klenov microscopic deterministic ATD model; the simulations have been performed on a single-lane road with an on-ramp bottleneck.

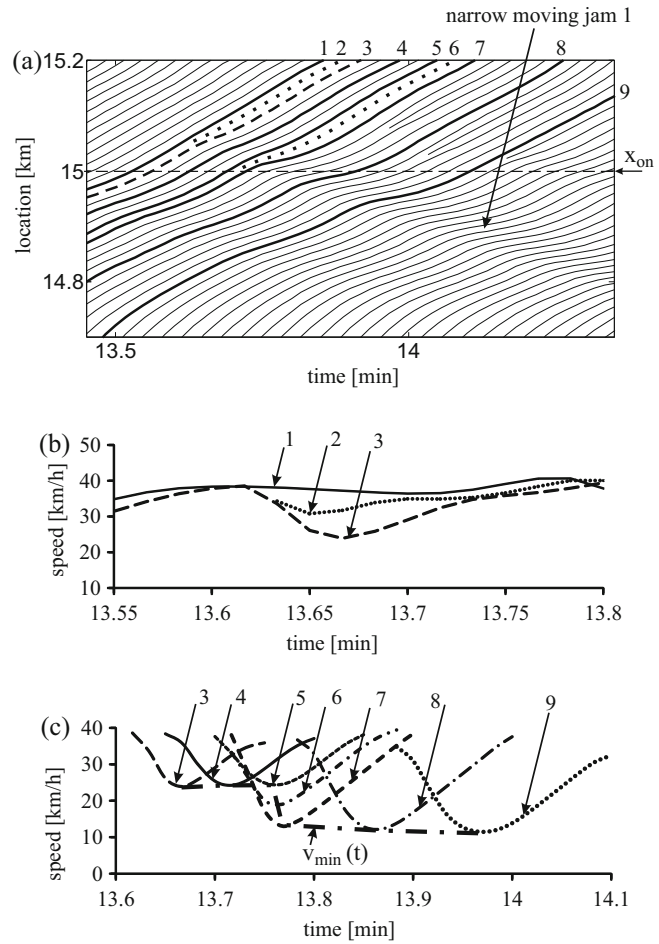
Some Scenarios of Development of Initial Speed Disturbances in Synchronized Flow

Firstly, we continue a study of the emergence of the narrow moving jam labeled by “narrow moving jam 1” in Figs. 21c and 24a. The initial speed disturbance initiating the emergence of this narrow moving jam is shown in Fig. 24b. The disturbance is caused by vehicle 2 merging from the on-ramp onto the main road (dotted vehicle trajectory 2 in Fig. 24a, b). A slower-moving vehicle 2 has no effect of the motion of the downstream-moving vehicle 1 (solid vehicle trajectory 1 in Fig. 24a, b). Contrarily, upstream vehicle 3 (dashed vehicle trajectory 3 in Fig. 24a, b) must decelerate strongly to avoid a collision with merging vehicle 2. As a result of this deceleration of vehicle 3, a disturbance of a local speed reduction appears. The minimum speed of vehicle 3 within this local disturbance is about $v_{\min} \approx 24$ km/h, whereas the speed of vehicle 1 is about 38 km/h.

The subsequent development of this initial speed disturbance is shown in Fig. 24c. We see that although all upstream-moving vehicles should decelerate while approaching vehicle 3, firstly, the amplitude of the disturbance does not grow. Indeed, the amplitude of the disturbance even slightly decreases: the minimum speed of upstream vehicle 5 becomes slightly larger than the minimum speed of vehicle 3. The situation changes when vehicle 6 (dotted vehicle trajectory 6 in Fig. 24a) merges from the on-ramp onto the main road. This vehicle forces upstream vehicle 7 to decelerate. As a result, the minimum speed of vehicle 7 becomes smaller than the minimum speed of the disturbance on vehicle trajectory 3 (Fig. 24a, c). This local decrease in the speed of vehicle 7 leads to a growing wave of the local speed reduction – an $S \rightarrow J$ instability is realized (vehicle trajectories 7–9 in Fig. 24a, c). The $S \rightarrow J$ instability results in the emergence of the narrow moving jam studied in section “Pinch Effect: Narrow Moving Jam Emergence in Synchronized

Spatiotemporal Features of Traffic Congestion,

Fig. 24 Continuation of Fig. 21. Initial speed disturbance and the development of $S \rightarrow J$ instability leading to the emergence of narrow moving jam labeled by “narrow moving jam 1” in Fig. 21c. (a) Fragments of vehicle trajectories related to Fig. 21c. (b) Microscopic speeds along vehicle trajectories 1, 2, and 3 labeled by the same numbers in (a). (c) Microscopic speeds along vehicle trajectories 1–9 labeled by the same numbers in (a). Dashed-dotted curve $v_{\min}(t)$ shows a time dependence of the minimum speed within the disturbance



Flow” (Fig. 21). Thus, we can see that already in this simple example, a nucleus causing the $S \rightarrow J$ instability has a complex spatiotemporal behavior: the nucleus is associated with two independent events of vehicle merging from the on-ramp onto the main road occurring at two different time instants.

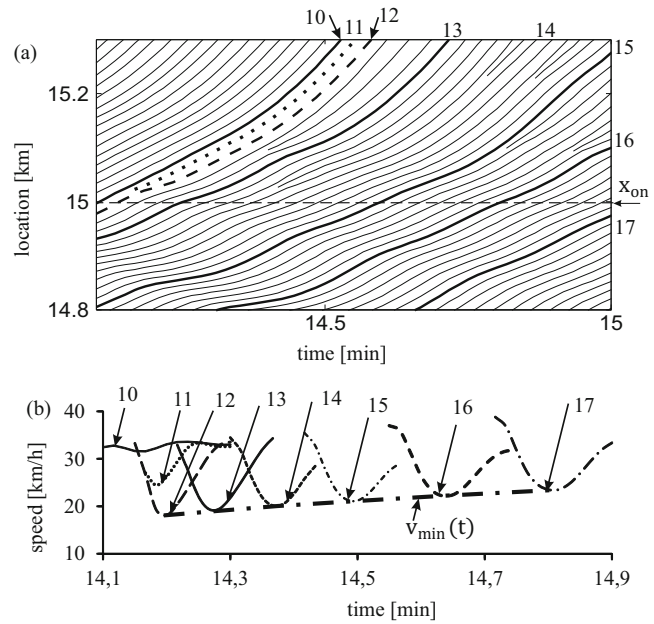
Almost any vehicle merging from the on-ramp onto the main road causes a disturbance of a local speed reduction in synchronized flow. Among such local disturbances, there are many small amplitude disturbances that decay over time. This is in accordance with the hypotheses about the nucleation nature of moving jam emergence in synchronized flow (Fig. 22) (Kerner 1998a, c, 1999a). However, even if the initial amplitude of a local disturbance is a large one, it can turn out that the driver speed adaptation overcomes the

driver over-deceleration in traffic flow upstream of the disturbance. As a result, the disturbance decays over time. This emphasizes once more that the hypothesis of section “[Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow](#)” that postulates features of hypothetical steady states of synchronized flow gives only a rough description of the nucleus emergence in synchronized flow. In reality, the emergence of a nucleus for the $S \rightarrow J$ instability is a complex spatiotemporal phenomenon that involves the behavior of many vehicles moving upstream of the initial disturbance.

In an example shown in Fig. 25, firstly, vehicle 11 merging from the on-ramp onto the main road (dotted vehicle trajectory 11 in Fig. 25a, b)

Spatiotemporal Features of Traffic Congestion,

Fig. 25 Continuation of Fig. 21. Decay of initial disturbance of local speed reduction in synchronized flow. (a) Fragments of vehicle trajectories related to Fig. 21c. (b) Microscopic speeds along vehicle trajectories 10–17 labeled by the same numbers in (a). Dashed-dotted curve $v_{\min}(t)$ shows a time dependence of the minimum speed within the disturbance



initiates a disturbance of a local speed reduction in synchronized flow: as in Fig. 24b, upstream vehicle 12 must decelerate strongly to avoid collision with vehicle 11. However, although the minimum speed $v_{\min} \approx 18$ km/h of vehicle 12 within the disturbance is about 6 km/h smaller than that of vehicle 3 in Fig. 24b, nevertheless, no S→J instability is realized: the initial speed disturbance decays over time (vehicle trajectories 13–17 in Fig. 25b). This emphasizes once more that the nucleus required for the S→J instability is a complex spatiotemporal phenomenon that depends both on the initial amplitude of a speed disturbance and on a spatiotemporal competition between the driver speed adaptation and over-deceleration effects occurring in synchronized flow upstream of the disturbance.

Effect of Initial Amplitude of Speed Disturbance in Pinch Region on Transformation of Growing Narrow Moving Jam into Wide Moving Jam

Another example of the effect of the amplitude of a disturbance of a local speed reduction in synchronized flow on the emergence of a nucleus for an S→J instability is shown in Figs. 26 and 27. In this case, the set of the flow rates q_{in} and q_{on} as well as other model parameters are the same as

those in Figs. 21, 24, and 25. The exclusion is a choice of a smaller parameter of vehicle merging at the on-ramp bottleneck. This leads to weaker safety conditions by vehicle merging from the on-ramp onto the main road than that used in Figs. 21, 24, and 25. Although due to these weaker safety conditions the mean time headway by vehicle merging decreases, nevertheless, no collisions between vehicles occur. The result of the weaker safety conditions by vehicle merging is a larger mean amplitude of initial local speed disturbances within the on-ramp merging region. We have found that at the same other model parameters as those in Figs. 21, 24, and 25, a larger initial local speed disturbance can appear: The minimum speed within the disturbance is about $v_{\min} \approx 7$ km/h (vehicle 3 in Fig. 26c, d).

The disturbance occurrence leads to the S→J instability with the emergence of a narrow moving jam (labeled by “narrow moving jam” in Fig. 26c). We can see that the microscopic criterion (2) for wide moving jams is not satisfied for the narrow moving jam: there is no flow interruption within the narrow moving jam (labeled by “no traffic interruption” in Fig. 27b).

The narrow moving jam grows continuously while transforming into a wide moving jam

(labeled by “wide moving jam” in Fig. 26b, c). This means that an S→J transition is realized (labeled by “S→J transition” in Fig. 26c). Indeed, there is a flow interruption interval $\tau_{\max} \approx 18.3$ s within the moving jam (labeled by “flow interruption interval is 18.3 s” in Fig. 27d) at road location $x = 14.65$ km. At parameters of the ATD model used for simulations, the mean time delay in vehicle acceleration at the downstream jam front of a wide moving jam is approximately equal to $\tau_{\text{del}}^{(\text{acc})}(0) \approx 1.8$ s. Thus, the microscopic criterion (2) for a wide moving jam is satisfied for the moving jam shown in Fig. 27c, d:

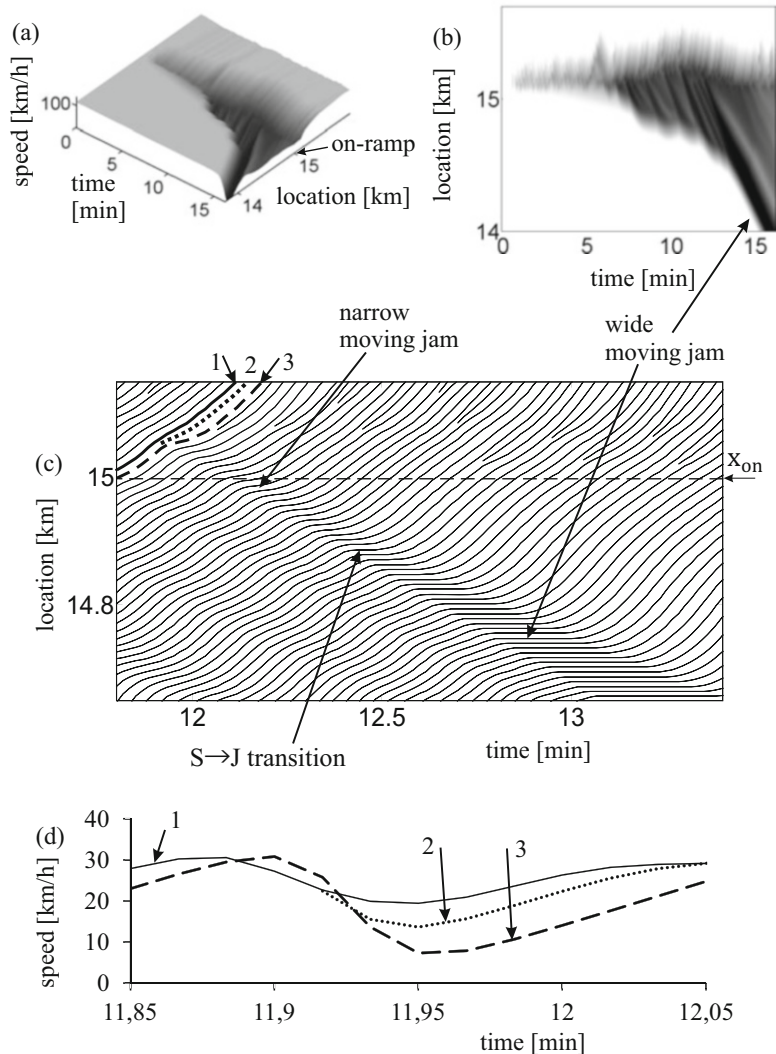
- A growing narrow moving jam can be considered a nucleus for the S→J transition.

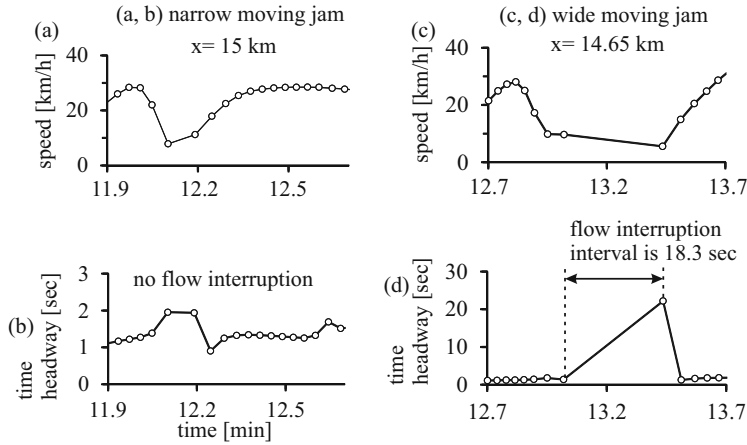
Disappearance of Pinch Effect as Consequence of Smaller On-Ramp Inflow

Now we decrease the on-ramp inflow rate from $q_{\text{on}} = 480$ vehicles/h used in simulations presented in Fig. 21 and Figs. 24–27 to $q_{\text{on}} = 400$ vehicles/h (Fig. 28), while all other model parameters of simulations shown in Fig. 21 remain the same as those in Fig. 21. The result of the reduction in the on-ramp inflow rate is an increase in the average speed of synchronized

Spatiotemporal Features of Traffic Congestion, Fig. 26

Simulations of the pinch effect with subsequent S→J transition. (a, b) Speed in space and time (a) and the same data presented by regions with variable shades of gray (shades of gray vary from white to black when the speed decreases from 90 km/h (white) to zero (black)) (b). (c) Fragments of vehicle trajectories. (d) Microscopic speeds along vehicle trajectories 1, 2, and 3 labeled by the same numbers in (c). Simulations of the Kerner-Klenov deterministic ATD model (Kerner and Klenov 2006) on a single-lane road with an on-ramp bottleneck at the same model parameters as those in Fig. 21 with the exclusion of parameter $\gamma = 0.15$ of the bottleneck model used in Kerner and Klenov (2006) (see Sect. 2.6 in Kerner and Klenov (2006)) that permits weaker safety conditions by vehicle merging at the on-ramp bottleneck





Spatiotemporal Features of Traffic Congestion, Fig. 27 Continuation of Fig. 26. (a, b) Microscopic vehicle speeds (a) and time headway (b) as time-functions measured through virtual detector at road location

$x = 15$ km in Fig. 26c. (c, d) Microscopic vehicle speeds (c) and time headway (d) as time-functions measured through virtual detector at road location $x = 14.65$ km in Fig. 26c

flow to $v_{\text{syn}} \approx 55$ km/h (Fig. 29a). This synchronized flow speed is considerably larger than the average speed in the pinch region $v_{\text{syn}} \approx 30$ km/h in Fig. 21 (measured through a virtual detector at the same location $x = 15$ km).

To explain this result, we should note that synchronized flow upstream of the bottleneck is maintained through the deceleration of vehicles moving on the main road within the on-ramp merging region. This vehicle deceleration is due to vehicles merging from the on-ramp onto the main road. The more frequent this vehicle deceleration is, the smaller the average synchronized flow speed. Therefore, the smaller the flow rate q_{on} , the smaller the frequency of the vehicle merging, and, therefore, the larger is the average speed of synchronized flow. Due to the larger synchronized flow speed, the mean minimum speed within local speed disturbances occurring within the merging region of the on-ramp in Fig. 28 is considerably larger than that in Fig. 24 (the minimum speed of vehicle 3 in Fig. 28e is $v_{\text{min}} \approx 41$ km/h, whereas the minimum speed of vehicle 3 in Fig. 24b is $v_{\text{min}} \approx 24$ km/h).

Because the minimum speed within local speed disturbances in synchronized flow within the merging region of the on-ramp in the case under consideration (Fig. 28e) is considerably larger than in the case shown in Fig. 24b, no

S→J instability and no narrow moving jams are realized in synchronized flow upstream of the on-ramp bottleneck (Fig. 28a–c). This result is in accordance with the hypothesis (e) of section “Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow.”

Spontaneous Emergence of Pinch Effect

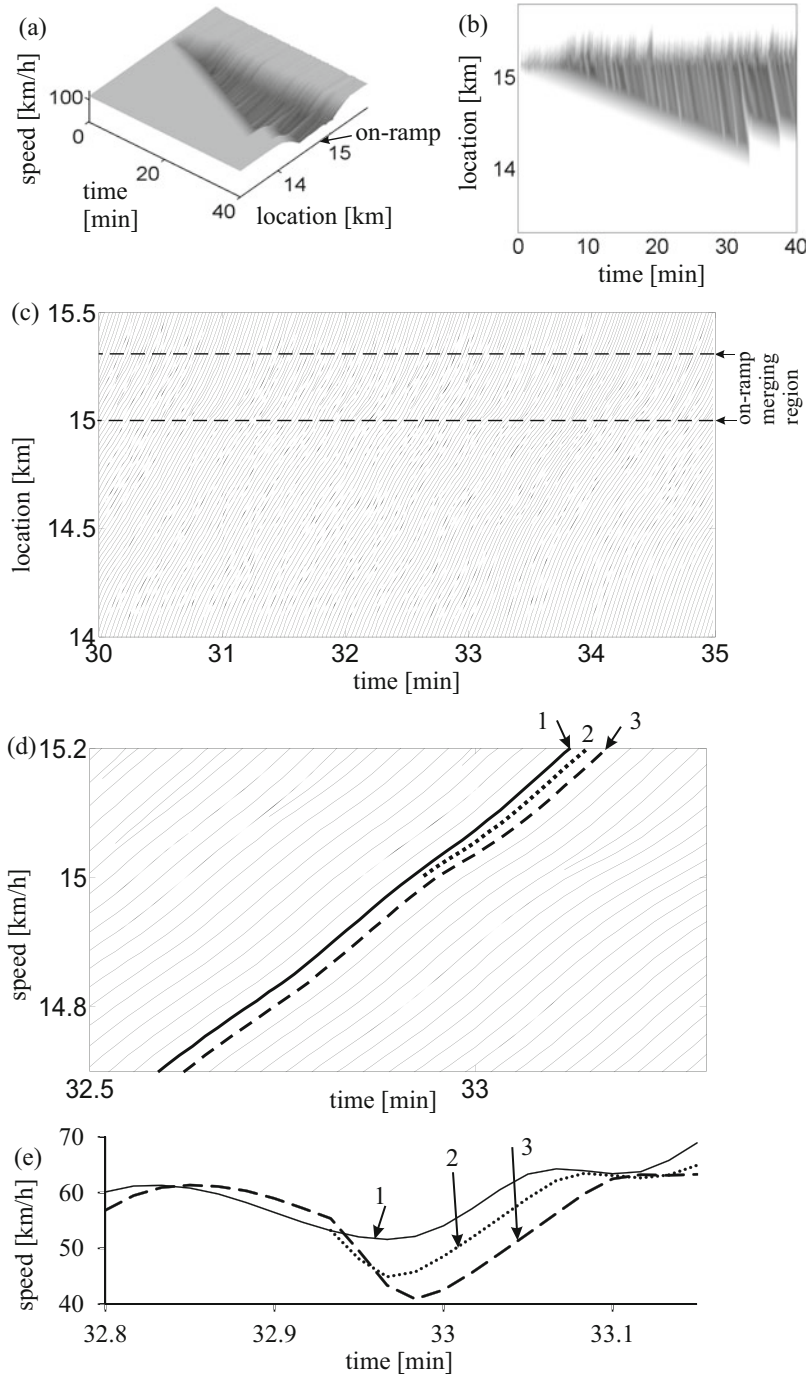
The effect of the on-ramp inflow reduction on the disappearance of the pinch effect in synchronized flow discussed above (Figs. 28 and 29) can exhibit qualitatively new features when we apply weaker safety conditions by vehicle merging from the on-ramp onto the main road (Figs. 30, 31, and 32). These weaker safety conditions are the same as those applied in Figs. 26 and 27. The result of the weaker safety conditions is a larger mean amplitude of local speed disturbances occurring due to vehicle merging from the on-ramp onto the main road within the on-ramp merging region.

We have found that during a long time interval, no S→J instability and no narrow moving jams are realized in synchronized flow upstream of the on-ramp bottleneck (time interval $0 < t < 33$ min in Fig. 30a–c). This is the same effect of the absent of the pinch region in synchronized flow as that shown in Fig. 28.

However, at time instant $t \approx 33$ min due to weaker safety conditions by vehicle merging

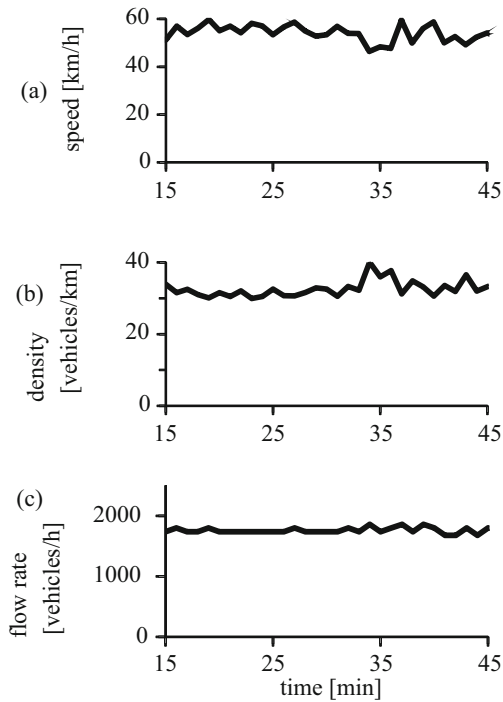
Spatiotemporal Features of Traffic Congestion,

Fig. 28 Simulations of synchronized flow at on-ramp bottleneck without occurrence of pinch effect. **(a, b)** Speed in space and time **(a)** and the same data presented by regions with variable shades of gray (shades of gray vary from white to black when the speed decreases from 90 km/h (white) to zero (black)) **(b)**. **(c, d)** Fragments of vehicle trajectories. **(e)** Microscopic speeds along vehicle trajectories 1, 2, and 3 labeled by the same numbers in **(d)**. Simulations of the Kerner-Klenov deterministic ATD model (Kerner and Klenov 2006) on a single-lane road with an on-ramp bottleneck at the same model parameters as those in Fig. 21 with the exclusion of smaller on-ramp inflow $q_{\text{on}} = 400$ vehicles/h



at the bottleneck, a large local speed disturbance appears: The minimum speed within the disturbance is about $v_{\text{min}} \approx 14$ km/h (vehicle trajectory 3 in Fig. 31). This minimum speed

is about 21 km/h smaller than the minimum speed of vehicle 3 in Fig. 28e when stronger safety conditions by vehicle merging have been used.



Spatiotemporal Features of Traffic Congestion, Fig. 29 Continuation of Fig. 28. (a–c) 1-min average vehicle speed (a), vehicle density (b), and flow rate (c) as time-functions measured through a virtual detector at road location $x = 15$ km in Fig. 28 (a–c)

The disturbance occurrence leads to the S→J instability with the emergence of a narrow moving jam (labeled by “narrow moving jam” in Fig. 31a). This means that the pinch effect is spontaneously realized in synchronized flow (the time instant of the occurrence of the pinch effect is symbolically shown by a dashed-dotted vertical line together with arrow labeled by “pinch” in Fig. 30b, c). In other words, the occurrence of a disturbance of a local speed reduction of a large enough amplitude in synchronized flow (vehicle trajectory 3 in Fig. 31) causes the occurrence of the pinch effect. Due to the pinch effect, the average speed in synchronized flow within the merging region of the on-ramp (drop in the average speed labeled by arrow “pinch” in Fig. 32a). This decrease in the average synchronized flow speed leads to the increase in the vehicle density within the pinch region of synchronized flow (jump in the density labeled

by arrow “pinch” in Fig. 32b). However, it should be noted that the emergence of the pinch effect in synchronized flow does not almost effect on the flow rate (Fig. 32c).

After the pinch effect has occurred, it is self-maintained in synchronized flow (arrow “pinch” in Figs. 30b, c and 32): a sequence of S→J instabilities with the subsequent emergence of several narrow moving jams is observed (labeled by “narrow moving jams” in Fig. 30d):

- The pinch effect can occur spontaneously in metastable synchronized flow in which no S→J instability has initially be realized. Such a pinch effect can be initiated by the spontaneous occurrence of a large-amplitude speed disturbance of a local speed reduction in synchronized flow.

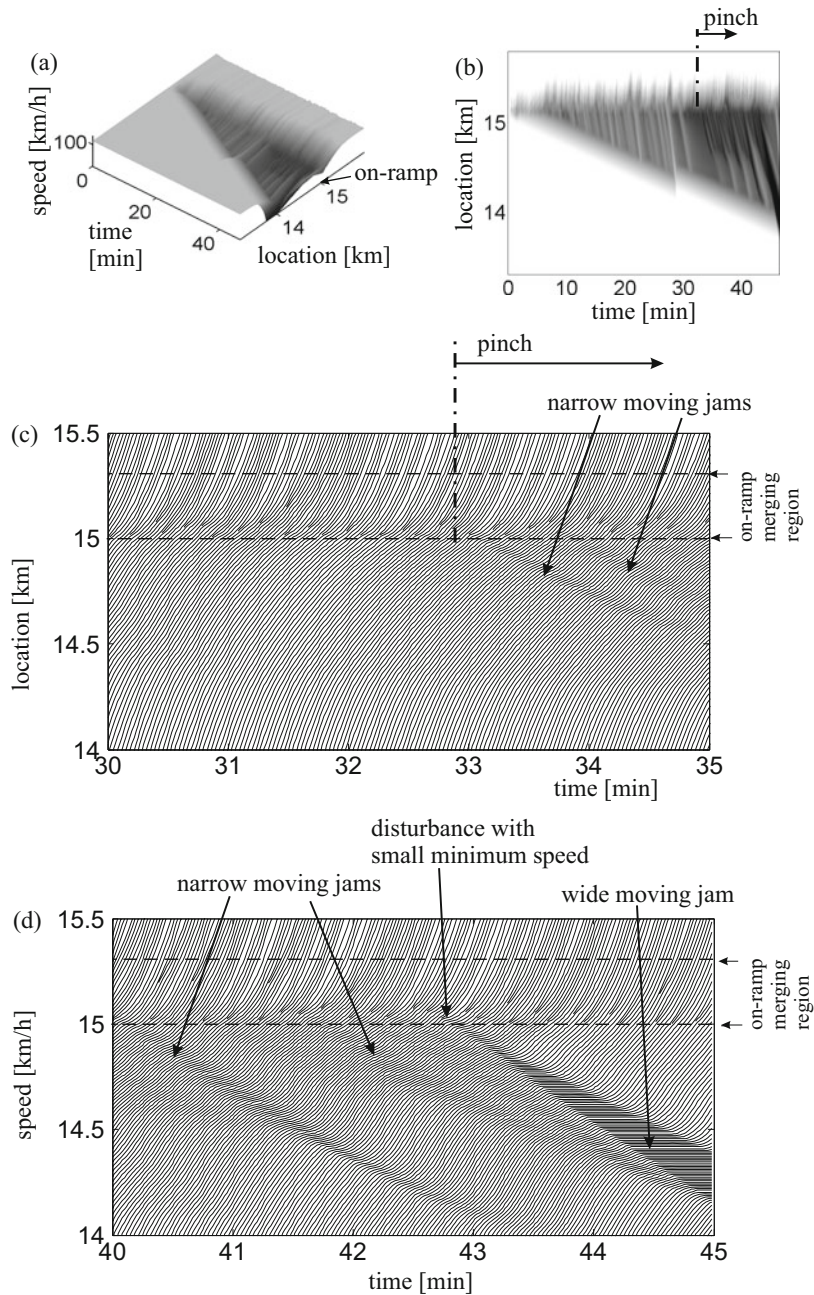
As in the case of the pinch effect shown in Fig. 26c, the pinch effect shown in Fig. 30 can lead to a narrow moving jam that subsequent growth can cause the emergence of a wide moving jam (labeled by “wide moving jam” in Fig. 30d). This S→J transition occurs, when a local disturbance of a large enough amplitude appears spontaneously in the pinch region of synchronized flow (labeled by “disturbance with small minimum speed” in Fig. 30d). The minimum speed within the disturbance $v_{\min} \approx 7.9$ km/h is almost as small as the minimum speed of vehicle 3 shown in Fig. 26c, d. This can explain why, as in Fig. 26c, in the case under consideration (Fig. 30d), the S→J transition is also realized.

Pinch Effect and Hypotheses of Three-Phase Theory About Moving Jam Emergence

In accordance with the hypotheses (c), (e), and (f) of the three-phase theory about moving jam emergence (section “[Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow](#)”), the pinch effect in synchronized flow should be characterized by a considerable deviation of metastable synchronized flow states that are above the line J in the flow–density plane from the line J (Figs. 22a and 33a) (Kerner 2004b). Indeed, the more the states of metastable synchronized flow deviate from the line J , the smaller the

Spatiotemporal Features of Traffic Congestion,

Fig. 30 Simulations of spontaneous occurrence of pinch effect. **(a, b)** Speed in space and time **(a)** and the same data presented by regions with variable shades of gray (shades of gray vary from white to black when the speed decreases from 90 km/h (white) to zero (black)) **(b, c, d)** Fragments of vehicle trajectories in two different time intervals. In **(b, c)**, dashed-dotted vertical lines together with arrows labeled by “pinch” show the time instant of spontaneous occurrence of pinch effect. Simulations of the Kerner-Klenov deterministic ATD model (Kerner and Klenov 2006) on a single-lane road with an on-ramp bottleneck at the same model parameters as those in Fig. 28 with the exclusion of parameter $\gamma = 0.15$ of the bottleneck model used in Kerner and Klenov (2006) that permits weaker safety conditions by vehicle merging at the on-ramp bottleneck



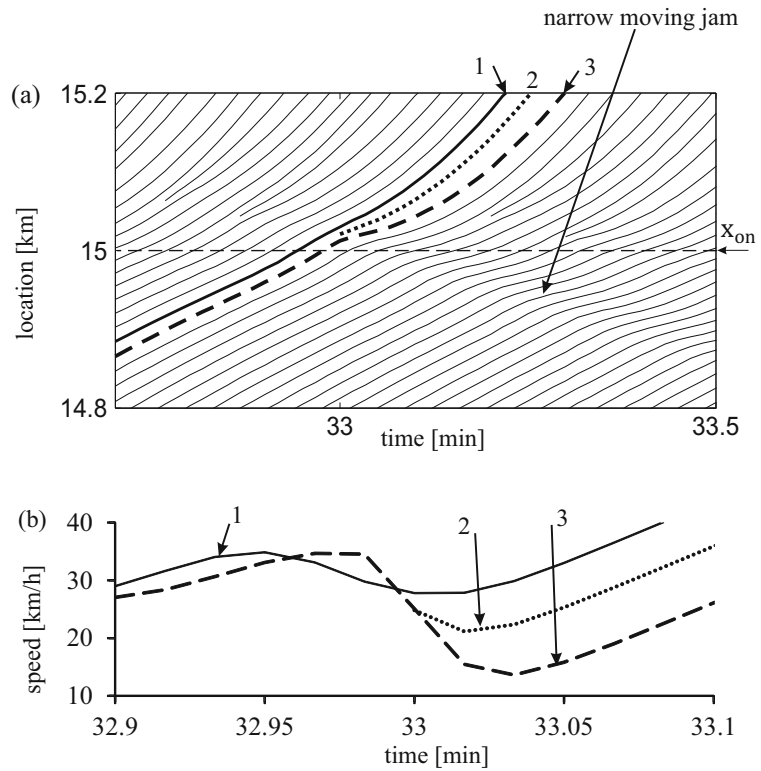
critical nucleus amplitude $\Delta v_{cr,SJ}$ required for an $S \rightarrow J$ instability (Fig. 22b) and, therefore, the larger the probability of the nucleus occurrence for the $S \rightarrow J$ instability. In accordance with these hypotheses of the three-phase theory, we have found that synchronized flow states related to the pinch effect shown in Fig. 21 deviate considerably from the line

J in the flow–density plane (diamond states labeled by “ S_{pinch} ” in Fig. 33b).

However, when the on-ramp inflow decreases and the pinch effect does not occur in synchronized flow at the on-ramp bottleneck as shown in Fig. 28, we can expect that synchronized flow states in the flow–density should come close to

Spatiotemporal Features of Traffic Congestion,

Fig. 31 Continuation of Fig. 30. (a) Fragment of vehicle trajectories shown in Fig. 30c. (b) Microscopic speeds along vehicle trajectories 1, 2, and 3 labeled by the same numbers in (a)



the line J . This effect is indeed realized in simulations (circle states labeled by “S” in Fig. 33b). Because the pinch effect does not occur in the case shown in Fig. 28, there is a density drop from the states “ S_{pinch} ” related to the pinch effect to states “S” in which no pinch effect occurs. The arrow labeled by “pinch disappearance” in Fig. 33b symbolizes this density drop caused by the disappearance of the pinch effect due to the decrease in the on-ramp inflow.

An opposite effect occurs in the case shown in Fig. 30. In this case, as explained above, the pinch effect occurs spontaneously at some time instant (arrows “pinch” in Fig. 30b, c). The pinch effect emergence is accompanied by a density jump (labeled by “pinch” in Fig. 34) from synchronized flow states labeled by “S” (no pinch effect is realized) to synchronized flow states labeled by “ S_{pinch} ” in Fig. 34 (pinch effect is realized). In the flow–density plane, synchronized flow states “ S_{pinch} ” related to the pinch effect deviate considerably from the line J .

S→J Instability in Driver Experiments of Car-Following

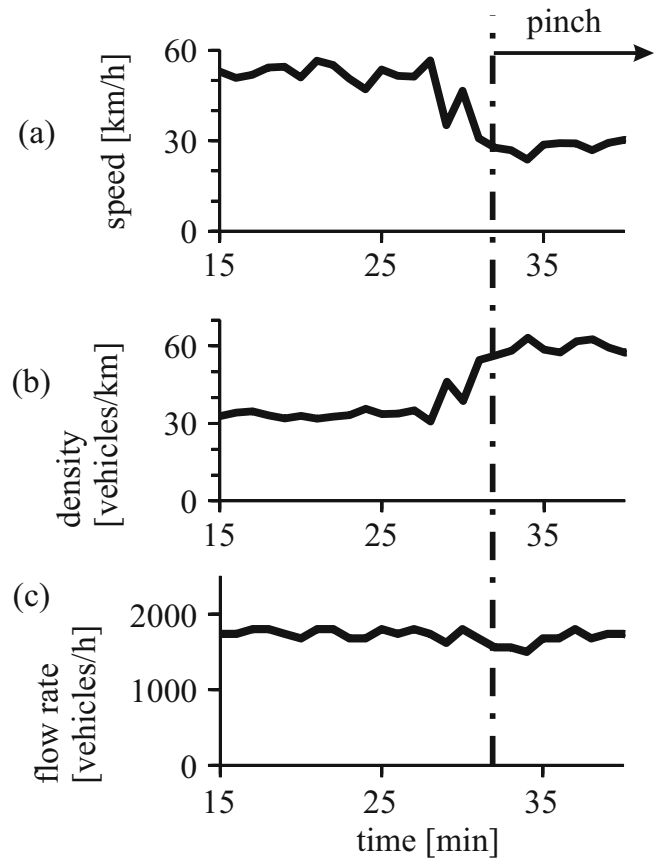
Experimental Moving Jam Emergence on Circular Road

Sugiyama et al. (2008), Nakayama et al. (2009), and Tadaki et al. (2013, 2016) have made driver experiments in which moving jams emerge in synchronized flow on a circular road without bottlenecks. For example, in Sugiyama et al. (2008) $N_{\text{veh}} = 22$ vehicles have moved along a single-lane circle road with $L_{\text{circle}} = 230$ m at the speed of about 30 km/h; the initial density in traffic flow has been $\rho = N_{\text{veh}}/L_{\text{circle}} \approx 95$ vehicles/km. Thus, the initial traffic flow in the experiment of Sugiyama et al. (2008), Nakayama et al. (2009), Tadaki et al. (2013, 2016) is congested traffic associated with the synchronized flow phase.

Simulations in the framework of the three-phase theory (Kerner 2008) (see Fig. 8.9 of the book Kerner (2017a)) made at approximately the

Spatiotemporal Features of Traffic Congestion,

Fig. 32 Continuation of Fig. 30. (a–c) 1-min average vehicle speed (a), vehicle density (b), and flow rate (c) as time-functions measured through a virtual detector at road location $x = 15$ km in Fig. 30 (c, d)



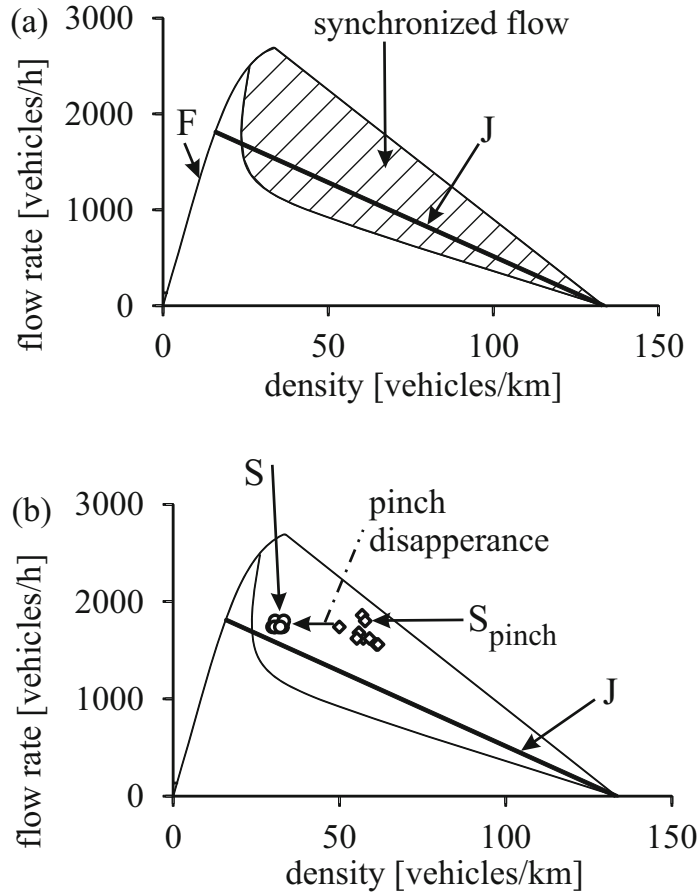
same parameters of synchronized flow as those used in real driver experiments of Sugiyama et al. (2008), Nakayama et al. (2009), Tadaki et al. (2013, 2016) have fully confirmed the hypotheses of the three-phase theory about the S→J instability and S→J transition discussed above (section “Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow”).

Experimental Concave Growth of Nuclei for S→J Instability

To study the development of an S→J instability, Jiang et al. (Jiang et al. 2014, 2015; Tian et al. 2016a, b) have performed driver experiment on an open section of a road. First, all vehicles have been in a standstill within a queue. Then, one after another, vehicles have escaped from the queue. The first (leading) vehicle has accelerated to a speed v_{lead} related to a synchronized flow speed (about or less than 40 km/h). While the leading vehicle has maintained the speed v_{lead} , all other

vehicles should not do it while following each other. In Jiang et al. (2014, 2017), speed disturbances of a large amplitude have appeared in synchronized flow caused by this car-following process. The disturbances have grown along the vehicle platoon. Jiang et al. (Jiang et al. 2014, 2015, 2017; Tian et al. 2016a, b) have found out that this disturbance growth occurs in a *concave* way along vehicles in the platoon leading to moving jam emergence. In Jiang et al. (2014), Jiang et al. (2015), Jiang et al. (2017), Tian et al. (2016a, b), it has been assumed that to simulate this concave growth pattern of oscillations, a *stochastic* traffic flow model is needed (Jiang et al. 2017, 2018; Tian et al. 2016c).

However, below we show that the concave growth pattern of oscillations found in Jiang et al. (2014, 2017) occurs also when simulations are made with a *deterministic* traffic flow model: no stochastic simulations of driver behavior and no applications of vehicle acceleration noise are needed to simulate the concave growth pattern of oscillations.



Spatiotemporal Features of Traffic Congestion, Fig. 33 Continuation of Figs. 21 and 28. Pinch effect and its disappearance in the flow–density plane: (a) Free flow (F), 2D steady states of synchronized flow (dashed region), and line J for the Kerner-Klenov deterministic ATD model (Kerner and Klenov 2006). (b) Disappearance of the pinch effect (labeled by arrow “pinch disappearance”); synchronized flow states (diamonds) labeled by “ S_{pinch} ” are related to time interval $13 \leq t \leq 21$ min in

Fig. 21a, b in which the pinch effect is realized at the on-ramp inflow flow rate $q_{\text{on}} = 480$ vehicles/h; synchronized flow states (circles) labeled by “S” are related to time interval $20 \leq t \leq 33$ min in Fig. 28a, b in which no pinch effect is realized at the on-ramp inflow flow rate $q_{\text{on}} = 400$ vehicles/h. 1-min average data is measured through the use of a virtual detector at $x = x_{\text{on}} = 15$ km. In (b), free flow states, 2D steady states of synchronized flow, and line J are taken from (a)

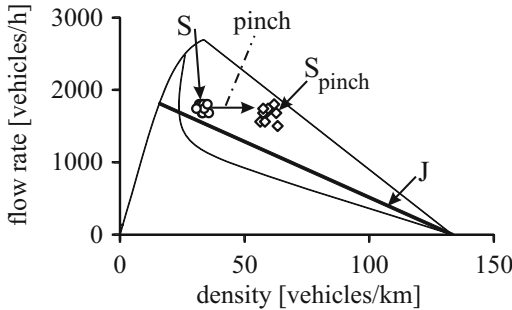
Simulations of Concave Growth of Nuclei for $S \rightarrow J$ Instability with Deterministic Traffic Flow Model
 Firstly, we consider the growth of speed disturbances in synchronized flow at the on-ramp bottleneck on the single-lane road shown in Fig. 21. These simulations have been made with the Kerner-Klenov deterministic ATD model. To find the standard deviation of the speed σ_i for the vehicle with the index i , we have used a well-known formula

$$\sigma_i = \sqrt{\frac{1}{N-1} \sum_{n=1}^N (v_i(t_n) - \bar{v}_i)^2}, \quad (6)$$

where $v_i(t_n)$ is the speed of vehicle i at time $t_n = n \Delta t + t_0$, $n = 1, 2, \dots, N$, $\Delta t = 0.1$ s, $N = (t_1 - t_0)/\Delta t$; t_0 and t_1 are the first and the last time moments for the standard deviation calculation, respectively. The first (leading) vehicle with $i = 1$ is chosen as the closest vehicle

upstream of road location $x = x_{\text{on}} = 15$ km at time t_n . A vehicle with index $i + 1$ is the following one for a vehicle with index i .

As discussed in section “Pinch Effect: Narrow Moving Jam Emergence in Synchronized Flow,” in the pinch region of synchronized flow at the on-ramp bottleneck, speed disturbances occurring within on-ramp merging region (nuclei for the $S \rightarrow J$ instability) grow leading to the

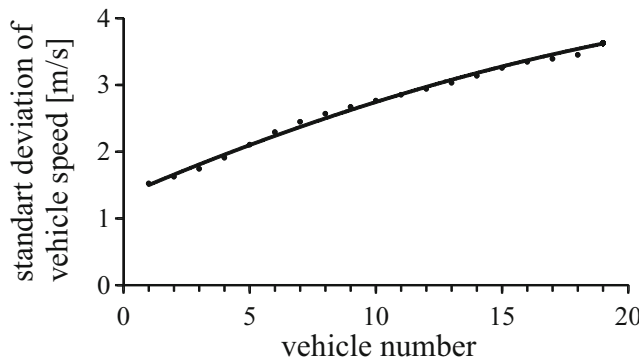


Spatiotemporal Features of Traffic Congestion, Fig. 34 Continuation of Fig. 30. Pinch effect emergence (arrow labeled by “pinch”) in the flow–density plane. Synchronized flow states (circles) labeled by “S” are related to synchronized flow at time interval $15 \leq t \leq 28$ min in which no pinch effect (no $S \rightarrow J$ instability) is realized. Synchronized flow states (diamonds) labeled by “S_{pinch}” are related to time interval $33 \leq t \leq 43$ min in which the pinch effect due to the $S \rightarrow J$ instability is realized. 1-min average data is measured through the use of a virtual detector at $x = x_{\text{on}} = 15$ km. Free flow states, 2D steady states of synchronized flow, and line J are taken from Fig. 33a

emergence of narrow moving jams (Fig. 21c). We have found that this growth of speed disturbances occurs in a concave way along vehicles in the platoon (Fig. 35). Therefore, the term “concave growth pattern of oscillations” used in Jiang et al. (2014, 2017) is equivalent to the term “concave growth of nuclei for the $S \rightarrow J$ instability” used in Fig. 35.

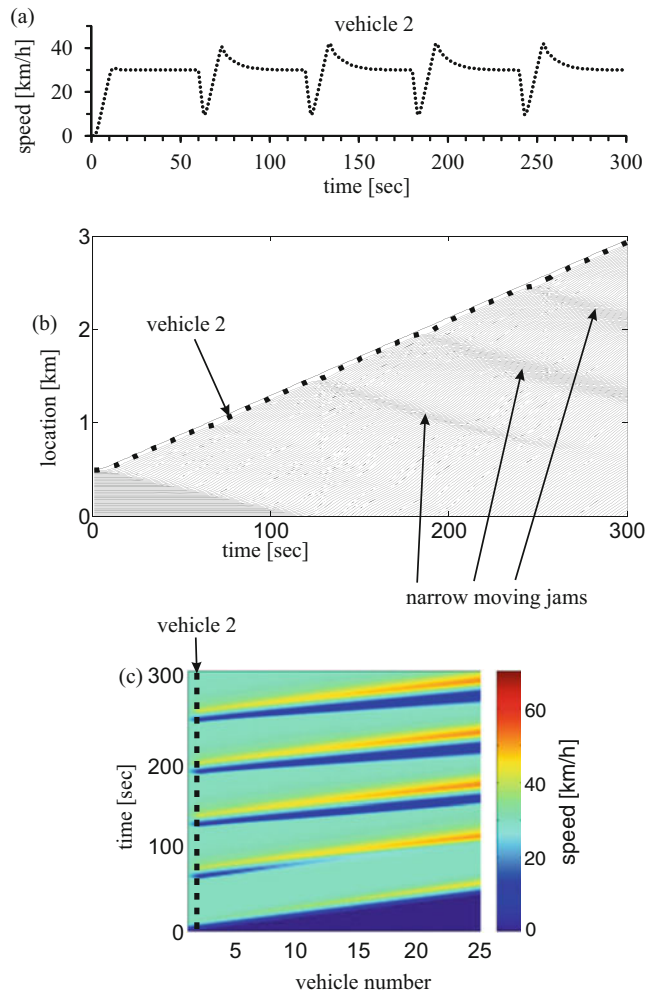
Simulations of Concave Growth of Speed Disturbances with Deterministic Model on Homogeneous Road

To simulate the concave growth of speed disturbances along vehicles in a vehicle platoon moving on a homogeneous road without bottlenecks with Kerner-Klenov deterministic microscopic ATD traffic flow model, we have induced a periodic (deterministic) large enough initial disturbance at the beginning of a vehicle platoon that follows the leading vehicle moving at a time-independent speed $v_{\text{lead}} = 30$ km/h: only vehicle 2 (dotted vehicle trajectory in Fig. 36a, b) that follows the leading vehicle produces a periodic short-time deceleration. Speed disturbances initiated by vehicle 2 begin to self-grow along the vehicle platoon (Fig. 36b, c). Similar to the concave way of the disturbance growth in synchronized flow at the on-ramp bottleneck (Fig. 35), the growth of the speed disturbances shown in Fig. 37 occurs in a concave way.



Spatiotemporal Features of Traffic Congestion, Fig. 35 Continuation of Fig. 21. Concave growth of nuclei for the $S \rightarrow J$ instability on a single-lane road with on-ramp bottleneck. Simulations of Kerner-Klenov

deterministic ATD model. Calculations of standard deviation of the speed are made with formula (6) during time interval $14 \leq t \leq 20$ min



Spatiotemporal Features of Traffic Congestion, Fig. 36 Simulations of speed disturbances with the Kerner-Klenov deterministic ATD model (Kerner and Klenov 2006) at the same model parameters as those given in Tables 1 and 2 of Kerner and Klenov (2006): pinch effect with the growth of speed disturbances. (a) Time dependence of speed of vehicle 2. (b) Vehicle trajectories. (c) Vehicle speed in space and time. The

periodic initial speed disturbance caused by vehicle 2 (dotted curves labeled by “vehicle 2” in (a–c)) has been simulated through deceleration of vehicle 2 with -1 m/s^2 ; after reaching the speed 10 km/h, vehicle 2 catches up the leading vehicle moving a constant speed $v_{\text{lead}} = 30 \text{ km/h}$, accordingly the ATD model rules of vehicle motion; this short-time deceleration of vehicle 2 has been repeated with period 1 min

In contrary to the pinch effect shown in Fig. 21, no pinch effect and, therefore, no growth of speed disturbances occur in synchronized flow at the on-ramp bottleneck in the case shown in Fig. 28. No pinch effect and no growth of speed disturbances occur also during the time interval $0 < t < 30 \text{ min}$ in Fig. 30. Both cases shown in Fig. 28 and at $0 < t < 30 \text{ min}$ in Fig. 30 are examples of a traffic situation in which synchronized flow

occurs in which speed disturbances are too small to create nuclei required for the S→J instability. A traffic situation, which is qualitatively similar to those shown in Fig. 28 and at $0 < t < 30 \text{ min}$ in Fig. 30 for the on-ramp bottleneck, is also realized in simulations of the queue discharge (Fig. 38). No large speed disturbances occur in this car-following process simulated with the deterministic ATD model. Therefore, no pinch effect can be

realized, and, therefore, no $S \rightarrow J$ instability occurs (Fig. 38). Respectively, as for synchronized flow at the on-ramp bottleneck (Fig. 28 and at $0 < t < 30$ min in Fig. 30) in which no $S \rightarrow J$ instability occurs (circles labeled by “S” in Figs. 33b and 34), synchronized flow state associated with Fig. 38 is related to a point in the flow–density plane that is only slightly above the line J (circle labeled by “S” in Fig. 39).

The opposite case is realized, when through periodic short-time deceleration of vehicle 2 a large initial disturbance has appeared in simulations (Figs. 36 and 37). In this case, the synchronized flow states in the flow–density plane labeled by “ S_{pinch} ” in Fig. 39 deviate considerably from the line J . Therefore, in accordance with the three-phase theory, the $S \rightarrow J$ instability with subsequent narrow moving jam emergence can easily occur in the synchronized flow states. In other words, narrow moving jam emergence shown in Fig. 36 can be considered the pinch effect. It should be noted that the pinch effect shown in Fig. 36 occurs in car-following process upstream of the leading vehicle moving at a time-independent speed. Contrarily, the pinch effect shown in Fig. 21 occurs in synchronized flow upstream of the on-ramp bottleneck.

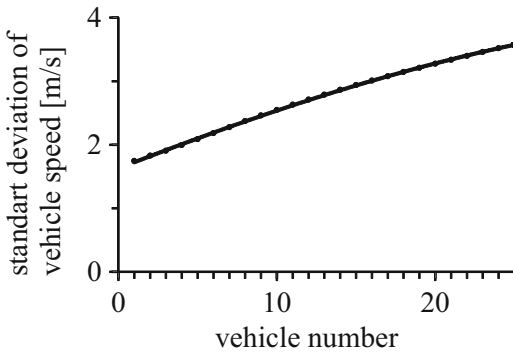
To explain the latter result in more details, we note that there is a density jump in the flow–density plane (arrow “pinch” in Fig. 39) from synchronized flow states labeled by “S” related to Fig. 38 to synchronized flow states labeled by “ S_{pinch} ” related to Fig. 36. The density jump in Fig. 39 has the same physical meaning as the density jump labeled by “pinch” in Fig. 34. As explained above, the density jump shown in Fig. 34 is caused by the pinch effect in synchronized flow at the on-ramp bottleneck. Therefore, as in synchronized flow at the on-ramp bottleneck (Fig. 34), the density jump in the flow–density plane (Fig. 39) is related to the pinch effect. Due to the pinch effect, synchronized flow states “ S_{pinch} ” deviate considerably from the line J . In accordance with the hypotheses of the three-phase theory (section “[Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow](#)”), in such synchronized flow

states (“ S_{pinch} ” in Fig. 39), the $S \rightarrow J$ instability with narrow moving jam emergence occurs spontaneously.

It should be noted that in Fig. 36 the leading vehicle moves at a constant speed; therefore, the density jump in the flow–density plane occurs at almost constant speed (labeled by arrow “pinch” in Fig. 39). Contrarily, synchronized flow at the bottleneck occurs at the flow rates that do not change; therefore, the density jump (labeled by arrow “pinch” in Fig. 34) occurs at almost constant value of the flow rate:

- As in synchronized flow at the bottleneck (Figs. 21, 30, and 34), in simulations of speed disturbances along vehicles in a vehicle platoon moving on a homogeneous road without bottlenecks with the Kerner-Klenov *deterministic* microscopic ATD traffic flow model it has been found that speed disturbances of *large amplitude* cause the pinch effect in synchronized flow (Figs. 36 and 39). The pinch effect results in the $S \rightarrow J$ instability with the concave growth of the speed disturbances (Fig. 37).
- The concave growth of the speed disturbances in synchronized flow is qualitatively the same on a single-lane road with the bottleneck (Fig. 35) and in a vehicle platoon moving on a homogeneous road without bottlenecks (Fig. 37). For this reason, the term “concave growth pattern of oscillations” used in Jiang et al. (2014, 2017) is equivalent to the term “concave way of the growth of nuclei for $S \rightarrow J$ instability.”
- The Kerner-Klenov deterministic microscopic ATD traffic flow model in the framework of the three-phase theory can describe concave growth pattern of oscillations as well as other empirical and experimental features of moving jam emergence in real traffic and driver experiments.
- In accordance with the hypotheses of the three-phase theory (section “[Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow](#)”), the concave growth of speed disturbances (oscillations) occurs in metastable synchronized flow with respect to

an $S \rightarrow J$ transition when a nucleus for the $S \rightarrow J$ instability appears in the synchronized flow. The concave growth of the nucleus can be explained by a spatiotemporal competition of the driver speed adaptation within 2D states of metastable synchronized flow introduced in the three-phase theory and the driver over-deceleration effect of Herman, Gazis, Montroll, Potts, Rothery, and Chandler (Gazis et al. 1961, 1959; Chandler et al. 1958; Herman et al. 1959).

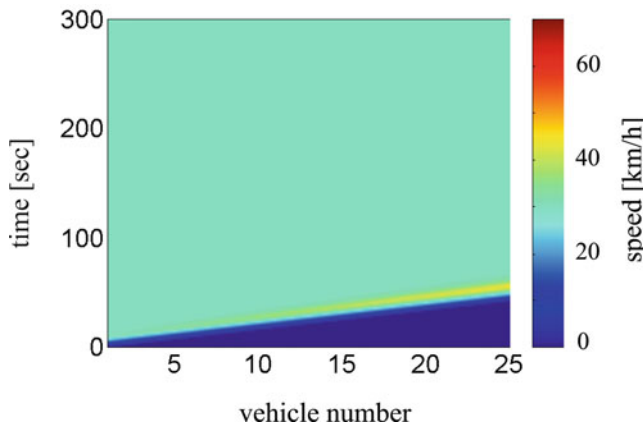


Spatiotemporal Features of Traffic Congestion, Fig. 37 Continuation of Fig. 36. Calculations of standard deviation of the vehicle speed: concave growth of nuclei for $S \rightarrow J$ instability

About Applicability of Results of Driver Experiments for Real Traffic

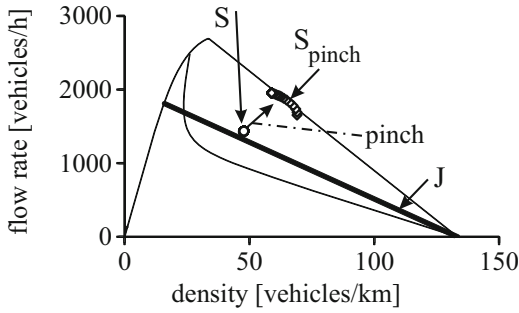
Driver experiments allow us to understand many important features of traffic flow that are difficult to study in real measured data. In particular, Sugiyama et al. driver experiment (Sugiyama et al. 2008; Nakayama et al. 2009; Tadaki et al. 2013, 2016) has proven the metastability of synchronized flow with result to an $S \rightarrow J$ transition. In Jiang et al. driver experiment, a concave growth of nuclei for an $S \rightarrow J$ instability has been found out (Jiang et al. 2014, 2017).

However, in driver experiments, the driver behavior can be strongly influenced by the objective of the experiment. Contrarily, in real traffic, each of the drivers decides itself how the driver should behave in a particular situation in car-following. For this reason, at least some of the results of driver experiments should be considered “artificial” results that have no relation to the reality. For example, if drivers move in real traffic on a single-lane road behind a slow-moving vehicle and the drivers see no chance to overtake it, the drivers move at large enough space gaps to each other at nearly the same speed as the speed of the slow-moving vehicle. In this case, in contrast with driver behavior in Jiang et al. driver experiment



Spatiotemporal Features of Traffic Congestion, Fig. 38 Simulations of vehicles following a leading vehicle moving at time-independent speed $v_{\text{lead}} = 30$ km/h after the vehicles have escaped from a queue of motionless vehicles. Simulations with the Kerner-Klenov

deterministic ATD model (Kerner and Klenov 2006). Model parameters are the same as those given in Tables 1 and 2 of Kerner and Klenov (2006). The result of simulations is similar to that presented in Tian et al. (2016c)



Spatiotemporal Features of Traffic Congestion, Fig. 39 Pinch effect in simulations of a vehicle platoon moving on a homogeneous road without bottlenecks shown in the flow–density plane. Synchronized flow state (circle) labeled by “S” is related to almost homogeneous synchronized flow shown in Fig. 38, in which drivers produce no large speed disturbances in car-following and, therefore, no pinch effect (no $S \rightarrow J$ instability) is realized. Synchronized flow states (diamonds) labeled by “ S_{pinch} ” are related to simulations shown in Figs. 36 and 37 in which vehicle 2 produces a periodic speed disturbance (with period 1 min) that results in the occurrence of the pinch effect. Data is measured through the use of moving averaging over 20 vehicles made within time interval 100–300 s. Free flow states, 2D steady states of synchronized flow, and line J are taken from Fig. 33a

(Jiang et al. 2014, 2017), no large speed disturbances would be produced by the drivers, and no growth of these disturbances would be observed.

The same drivers can exhibit a qualitatively different behavior in synchronized flow on a multilane road while the drivers approach the slow-moving vehicle. In this case, the drivers try to change the lane and to pass the slow-moving vehicle. Through the lane changing, large-amplitude speed disturbances can appear in the synchronized flow resulting in the occurrence of an $S \rightarrow J$ instability.

Nucleation-Interruption Effect Resulting in Alternations of Jam Emergence and Dissolution in Synchronized Flow

Definition of Nucleation-Interruption Effect Associated with Moving Jam Emergence

In addition to spatiotemporal phenomena of the nucleation of an $S \rightarrow J$ instability discussed above, there can be a nucleation-interruption effect associated with moving jam emergence that is as

follows. Moving jam nucleation in synchronized flow ($S \rightarrow J$ instability), i.e., the growth of an initial local speed disturbance, can be interrupted (see Sect. 6.5.3 of the book Kerner 2004b). This interruption of moving jam nucleation can occur if a space gap of one of the drivers that reaches a growing narrow moving jam is considerably larger than the mean space gap within this disturbance. In this case, this driver will not necessarily decelerate stronger than the preceding vehicle: this driver has enough time to adjust the speed to that of the preceding vehicle. As a result, the development of the $S \rightarrow J$ instability can be interrupted:

- The nucleation-interruption effect associated with moving jam emergence is a sequence of the emergence and dissolution of a narrow moving jam in synchronized flow.

The nucleation-interruption effect in synchronized flow can be responsible for a complex spatiotemporal process in which the emergence of growing narrow moving jams (growing nuclei for $S \rightarrow J$ transitions) alternates with the jam dissolution.

About Moving Jam Emergence on Multilane Roads
A random sequence of nucleation-interruption effects can be often observed in synchronized flow on multilane highways at which there are several adjacent on- and off-ramps located closely to each other. Indeed, due to lane changing, spatiotemporal phenomena of the nucleation, development, and interruption of the $S \rightarrow J$ instability can be considerably more complex on a multilane road than those on a single-lane road.

As found in simulations of moving jam emergence in synchronized flow on multilane roads in the framework of the three-phase theory (Kerner and Klenov 2009), random spatiotemporal nucleation-interruption effects determine the moving jam dynamics (called often as traffic oscillations) observed in a well-known NGSIM-single-vehicle data (Next Generation Simulation Programs 2016). Indeed, due to vehicle merging from the on-ramp and the subsequent lane changing, speed disturbances of large amplitudes can

easily occur. These speed disturbances can be nuclei for the $S \rightarrow J$ instability. Contrarily, at an off-ramp, vehicles leaving the main road can cause large space gaps between vehicles remaining on the main road. This can interrupt the nucleus growth, i.e., the nucleation-interruption effect can be realized leading to jam dissolution. Therefore, if the on- and off-ramps are close to each other (less than 1–2 km), as this is the case in NGSIM-single-vehicle data, growing narrow moving jams can dissolve before the jams transform into wide moving jams. In addition to Kerner and Klenov (2009), the empirical moving jam dynamics observed in NGSIM-single-vehicle data has been studied in many other works (see, e.g., Chen et al. (2012a, b), Laval et al. (2014) and references there).

It has been shown (Kerner 2009c; Kerner and Klenov 2009) that all known diverse spatiotemporal phenomena of the nucleation, development, and interruption of the $S \rightarrow J$ instability on multi-lane roads can be qualitatively explained through the hypotheses of moving jam emergence of the three-phase theory discussed in section “[Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow](#).” A detailed analysis of complex nucleation-interruption effects in synchronized flow due to lane changing has been made in Kerner and Klenov (2009) (see also Sect. 5.3 and explanations to Fig. 7.25 of the book Kerner 2009c).

Congested Patterns at Road Bottlenecks

Although traffic breakdown ($F \rightarrow S$ transition) can occur in free flow outside bottlenecks, in the most cases of the occurrence of traffic congestion in real traffic, an initial $F \rightarrow S$ transition leading to traffic congestion occurs at a bottleneck (in particular, road bottlenecks that are caused by on- and off-ramps). For this reason, in this review article, we restrict a consideration of congested traffic patterns that have appeared due to traffic breakdown at the bottleneck. Features of congested patterns occurring on a homogeneous road can be found in Chap. 17 of the book Kerner (2004b) as well as in the review article Kerner (2009b).

In addition to road bottlenecks, there are moving bottlenecks in traffic flow caused by slow-moving vehicles. As found in real measured traffic data, moving bottlenecks effect considerably on features of traffic breakdown ($F \rightarrow S$ transition) at road bottlenecks (Kerner et al. 2015). An $F \rightarrow S$ transition can also occur at moving bottlenecks outside road bottlenecks. Such an $F \rightarrow S$ transition can lead to the occurrence of a congested pattern that downstream front propagates downstream with the speed of the moving bottleneck. However, spatiotemporal features of congested patterns occurring at moving bottlenecks whose theory has been derived by Kerner and Klenov (2010) are out of scope of this paper.

A road bottleneck at which traffic breakdown occurs is called the *effectual bottleneck*. After traffic breakdown has occurred, synchronized flow is formed at the effectual bottleneck. In many cases, the downstream front of the synchronized flow is fixed at the bottleneck (the exclusion is moving synchronized flow patterns). The location in a neighborhood of the bottleneck at which this downstream front of synchronized flow is fixed is called an *effective location of the effectual bottleneck* (the effective location of the bottleneck for short). It must be noted that the effective location of the bottleneck can be different from the location at which traffic breakdown ($F \rightarrow S$ transition) has occurred leading to congested pattern emergence. Moreover, both the location of traffic breakdown and the effective location of the bottleneck are probabilistic values in real traffic. Even for the same type of congested pattern, the effective location of the bottleneck can randomly change over time (Kerner 2004b).

On real freeways there are many effectual bottlenecks. When a congested pattern has occurred at a downstream effectual bottleneck, the pattern can propagate further upstream reaching an upstream effectual bottleneck. In this case, if free flow at the upstream bottleneck is in a metastable state with respect to an $F \rightarrow S$ transition (i.e., with respect to synchronized flow emergence), then congestion pattern propagation from the downstream bottleneck causes an induced $F \rightarrow S$ transition at the upstream bottleneck. This can lead to complex congested pattern transformation effects considered below.

Before we discuss these complex effects of congested pattern transformation (see sections “[Complex Congested Patterns and Pattern Interaction at Adjacent Bottlenecks](#)” and “[Reproducible and Predictable Congested Patterns](#)”), in sections “[Synchronized Flow Patterns](#),” “[General Congested Patterns](#),” “[Heavy Traffic Congestion](#),” “[The Fundamental Requirement for Reliability of Intelligent Transportation Systems \(ITS\)](#),” and “[Diagram of Congested Patterns at Isolated Bottlenecks](#),” we would like firstly to discuss congested patterns that emerge at a hypothetical isolated effectual bottleneck. As we have already mentioned, on real freeways, there are many effectual bottlenecks; therefore, a consideration of the isolated bottleneck is an idealization of real traffic. Obviously, this idealization is easy to perform with a traffic flow model rather than with real traffic. However, on real freeways, there can be freeway sections with adjacent effectual bottlenecks that are located far enough from each other. On such freeway sections during long enough time intervals before a congested pattern, which has emerged at a downstream bottleneck, reaches an upstream bottleneck, the congested pattern can be considered the pattern at an isolated bottleneck. This bottleneck is the downstream bottleneck at which the pattern has initially occurred. Such a simplification of the reality allows us to understand some important common spatiotemporal features of congested patterns. Nevertheless, although in sections “[Synchronized Flow Patterns](#),” “[General Congested Patterns](#),” “[Heavy Traffic Congestion](#),” “[The Fundamental Requirement for Reliability of Intelligent Transportation Systems \(ITS\)](#),” and “[Diagram of Congested Patterns at Isolated Bottlenecks](#)” only spatiotemporal features of congested patterns at isolated bottlenecks are discussed, all illustrations of *empirical* examples of congested patterns represent the real development of measured congested patterns, which sometimes cause induced traffic breakdowns and complex pattern transformations at upstream bottlenecks when the patterns reach them. These real induced pattern formation and complex congested pattern transformation will be further explained in sections “[Complex Congested Patterns and Pattern](#)

[Interaction at Adjacent Bottlenecks](#)” and “[Reproducible and Predictable Congested Patterns](#).”

There are two main types of congested patterns at an isolated bottleneck:

1. Synchronized flow patterns (SP for short). An SP is a congested traffic pattern that consists of the synchronized flow phase only.
2. General patterns (GP for short). A GP is a congested traffic pattern that consists of both the synchronized flow and wide moving jam phases.

Synchronized Flow Patterns

As a result of an F→S transition at an isolated bottleneck, various SPs can occur at the bottleneck. There are three main types of SPs on highways (Kerner 2004b):

- (i) Localized SP (LSP for short): The downstream front of an LSP is fixed at the bottleneck. The upstream front of the LSP does not continuously propagate upstream over time: this front is localized at some distance upstream of the bottleneck. However, the location of the upstream front of synchronized flow in the LSP and therefore the LSP width (in the longitudinal direction) can exhibit complex oscillations over time. Note that the upstream front of synchronized flow separates free flow upstream from synchronized flow downstream of this front. Vehicles must slow down within the upstream front of synchronized flow.
- (ii) Widening SP (WSP for short): As in the LSP, the downstream front of a WSP is fixed at the bottleneck. In contrast to the LSP, the upstream front of the WSP propagates upstream continuously over time. In other words, the width of synchronized flow (in the longitudinal direction) in the WSP widens in the upstream direction. It must be noted that due to this widening, the upstream front of synchronized flow must eventually reach an upstream adjacent effectual bottleneck. It can turn out that at this upstream bottleneck, another congested pattern either already exists or can be induced. In both

cases, the WSP can be caught at this upstream bottleneck (catch effect; see below). Thus, in real traffic, the WSP can usually exist only for a finite time, before the WSP reaches the upstream effectual bottleneck.

- (iii) Moving SP (MSP for short): In contrast to the LSP and WSP, an MSP is a localized pattern on a road that propagates over time between road bottlenecks. The MSP is spatially limited by downstream and upstream MSP fronts. When an MSP propagates through free flow states, then depending on the flow rate within the MSP and the flow rate in free flow outside the MSP, each of the MSP fronts can propagate either upstream or downstream. In particular, downstream MSP propagation (i.e., when both MSP fronts propagate downstream) is possible; this occurs if the flow rate within the MSP is larger than the flow rates in free flows outside the MSP. Otherwise, upstream MSP propagation is realized. Due to MSP propagation, the MSP (or one of the MSP fronts) can reach an adjacent effectual bottleneck at which synchronized flow can be induced; then, the MSP is caught at the bottleneck (catch effect). Thus, the MSP that propagates between bottlenecks can exist only for a finite time, before the MSP (or one of the MSP fronts) reaches a nearest adjacent

effectual bottleneck at which synchronized flow can be induced.

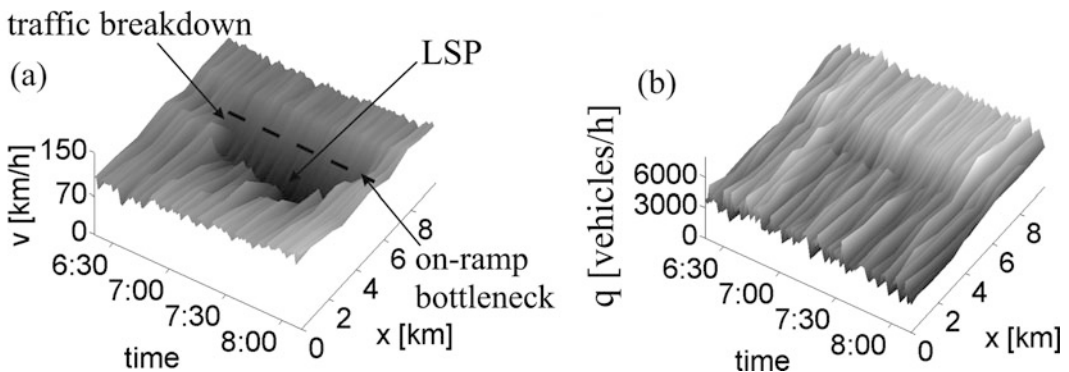
Localized SP (LSP)

An empirical example of an LSP is shown in Fig. 40. It can be seen that there are no wide moving jams in the LSP. Moreover, we can see that whereas the speed is considerably lower within the LSP (Fig. 40a), the flow rate is close to the flow rate in free flow (Fig. 40b, where the increase in flow rate downstream of the bottleneck is related to the on-ramp inflow).

The upstream front of synchronized flow in the LSP (Fig. 40a) does not continuously propagate upstream of the bottleneck over time: this upstream front is localized at some finite distance upstream of the bottleneck. This distance is a function of time.

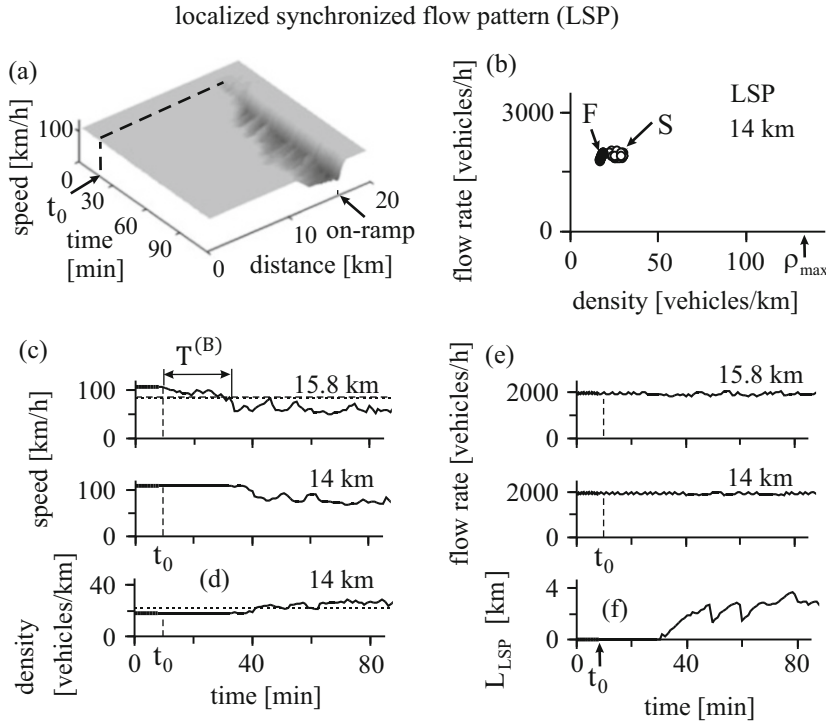
In numerical simulations made in the framework of the three-phase theory, LSPs exhibit qualitatively the same spatiotemporal features as those found in empirical data (Fig. 41) (Kerner and Klenov 2003; Kerner 2004b):

- (i) The downstream front of the LSP is fixed at the on-ramp bottleneck. However, the upstream front of the LSP is localized on the main road at some distance L_{LSP} upstream of the on-ramp bottleneck (Fig. 41a, f).
- (ii) Whereas the speed within the LSP is considerably lower than in free flow, the flow rate



Spatiotemporal Features of Traffic Congestion, Fig. 40 Empirical example of a localized synchronized flow pattern (LSP) at on-ramp bottleneck. (a) Averaged vehicle speed in the LSP in space and time. (b) Total flow

rate across the freeway in space and time. The dashed line in (a) shows the location of the bottleneck. (Adapted from Kerner 2004b)



Spatiotemporal Features of Traffic Congestion, Fig. 41 Simulated localized synchronized flow pattern (LSP) at on-ramp bottleneck in the framework of the three-phase theory (Kerner and Klenov 2003): (a) Averaged vehicle speed in the LSP in space and time. (b) Free flow (F) and synchronized flow within the LSP (S) in the flow-density plane. (c–e) Speed (c), density (d), and flow

rate (e) as time-functions at fixed freeway locations. (f) The LSP width L_{LSP} as a time-function. t_0 is a time instant at which the on-ramp inflow has been switched on in simulations. In (c), $T^{(B)}$ is a time delay of traffic breakdown at the bottleneck leading to LSP formation. The beginning of the on-ramp merging region is at $x_{on} = 16$ km. (Adapted from Kerner 2004b)

is approximately the same as that in free flow (Fig. 41b–e).

- (iii) The LSP width L_{LSP} (in the longitudinal direction) can be a random time-function (Fig. 41f). The mean width of the LSP $L_{LSP}^{(mean)}$ can depend strongly on q_{in} and q_{on} . In numerical simulations, different LSPs have been found for which $L_{LSP}^{(mean)}$ varies between 0.5 and 10 km.

Widening SP (WSP)

An empirical example of a WSP is shown in Fig. 42. In this case, synchronized flow is formed due to a spontaneous traffic breakdown at an off-ramp bottleneck (labeled by “off-ramp bottleneck” in Fig. 42). The downstream front of this synchronized flow is fixed at this bottleneck. The upstream front of synchronized flow propagates

continuously upstream over time. When this front reaches an upstream on-ramp bottleneck (labeled by “on-ramp bottleneck” in Fig. 42), the WSP is caught at the bottleneck (catch effect) causing an induced traffic breakdown at the bottleneck with the subsequent formation of another congested pattern upstream of the bottleneck.

Numerical simulations of WSPs made in the framework of the three-phase theory show qualitatively the same spatiotemporal features as those found in empirical data (Fig. 43) (Kerner 2004b). After a random time delay $T^{(B)}$ of traffic breakdown, the WSP occurs spontaneously on the main road at the off-ramp (Fig. 43a) or on-ramp (Fig. 43b) bottlenecks. The downstream front of a WSP is fixed at the on-ramp bottleneck. The upstream front of the WSP is continuously widening upstream. As within other SPs, the speed

within the WSP is considerably lower than in free flow, and the flow rate is approximately the same as that in free flow (Figs. 42b and 43c, d).

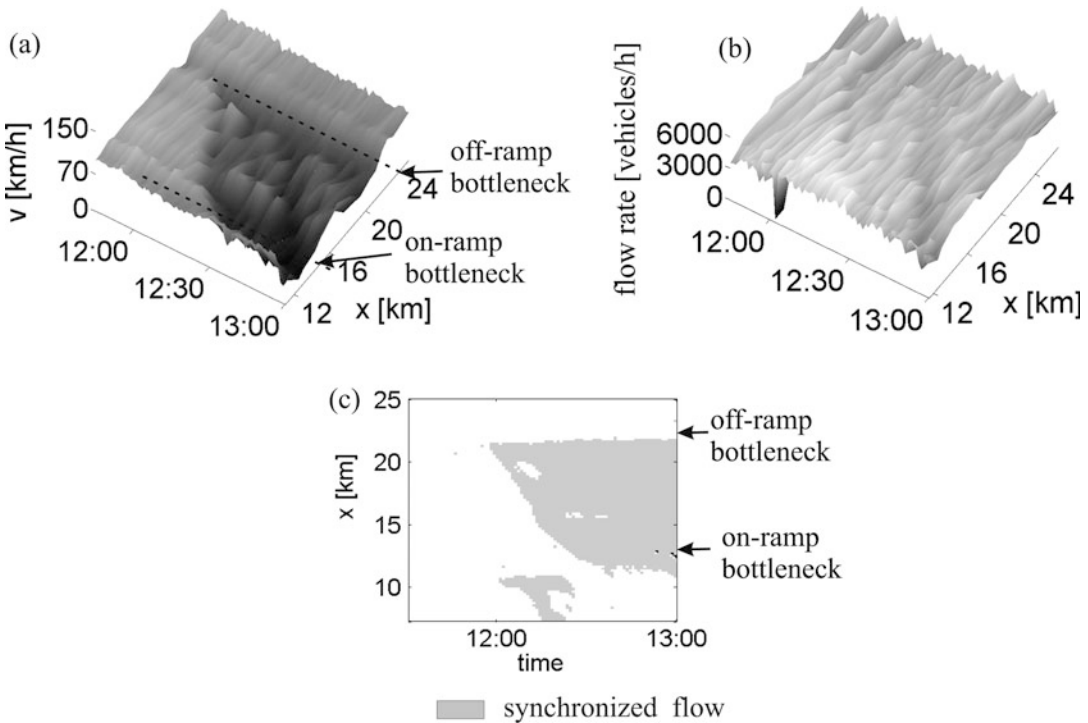
Moving SP (MSP)

An empirical example of an MSP is shown in Figs. 44 and 45. We consider two adjacent bottlenecks, a downstream off-ramp bottleneck (labeled by “off-ramp bottleneck”) and an upstream bottleneck (labeled by “on-ramp bottleneck 1”). Synchronized flow occurs spontaneously at the downstream off-ramp bottleneck (this traffic breakdown is labeled “spontaneous $F \rightarrow S$ transition” in Fig. 44c). The synchronized flow subsequently propagates upstream as a distinct localized structure. This is an example of an MSP (MSP is labeled “MSP” in Fig. 44a, c). When the MSP reaches the upstream bottleneck (on-ramp bottleneck 1), it does not propagate through the bottleneck as a localized structure: the MSP is caught at the bottleneck (catch effect).

This is in contrast to wide moving jam propagation through an upstream bottleneck, as shown in Fig. 1, and for a wide moving jam labeled “wide moving jam” in Fig. 44b, c. It should be also noted that whereas the flow rate is very small within the latter wide moving jam, we cannot almost distinguish the MSP in the flow rate distribution in space and time shown in Fig. 44b. As already mentioned above, this is a very important feature of many empirical (measured) SPs: the flow rate within the SP can be as large as in an initial free flow (Fig. 45a, b).

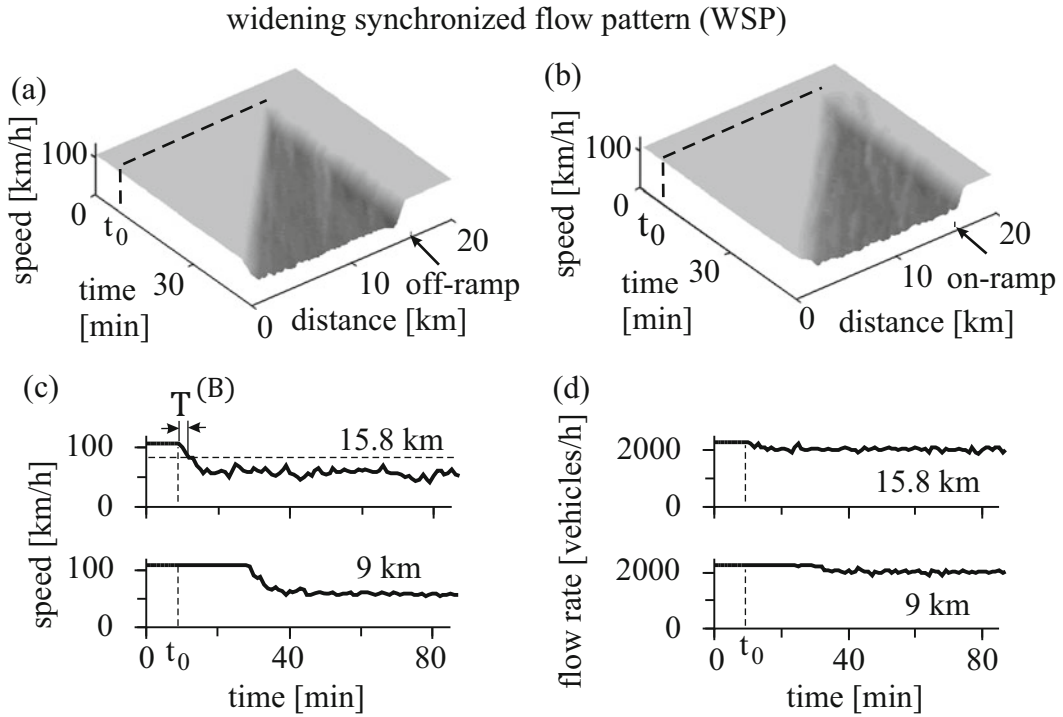
Numerical simulations of MSPs made in the framework of the three-phase theory show qualitatively the same spatiotemporal features as those found in empirical data (Figs. 44 and 45a):

- (i) An MSP occurs spontaneously on the main road either at the off-ramp (Fig. 46a) or on-ramp (Fig. 46b) bottlenecks. In each of these two cases, after traffic breakdown has



Spatiotemporal Features of Traffic Congestion, Fig. 42 Empirical example of a WSP and the catch effect: (a, b) Average speed (a) and flow rate (b) in space and

time. (c) Graph of (a) with overview of free flow (white) and synchronized flow (gray). (Adapted from Kerner 2004b)



Spatiotemporal Features of Traffic Congestion, Fig. 43 Simulated widening synchronized flow patterns (WSP) at off-ramp (a) and on-ramp (b–d) bottlenecks in the framework of three-phase theory (Kerner and Klenov 2003): (a, b) Averaged vehicle speed in the WSP in space and time. (c, d) Speed (c) and flow rate (d) as time-functions at a fixed freeway location related to (b). t_0 is a

time instant at which the off-ramp outflow (a) and the on-ramp inflow (b–d) have been switched on in simulations. In (c), $T^{(B)}$ is a time delay of traffic breakdown at the bottleneck leading to WSP formation. In (a), the beginning of the off-ramp lane is at $x_{\text{off}} = 16$ km. In (b–d), the beginning of the on-ramp merging region is at $x_{\text{on}} = 16$ km. (Adapted from Kerner 2004b)

occurred at the bottleneck, the downstream front of synchronized flow departs from the bottleneck. As a result, the MSP begins to propagate on the main road as an independent localized structure on a homogeneous road (Fig. 46a, b). As in empirical observations (Fig. 45a, b), whereas the speed within the MSP is considerably lower than in free flow (Fig. 46c), the flow rate is approximately the same as that in free flow (Fig. 46d).

- (ii) Simulations show (see the book Kerner (2004b)) that when two adjacent bottlenecks exist on a freeway and an MSP appears at the downstream one, then as in empirical data (Fig. 44a, c), the MSP can be caught at the downstream bottleneck with the subsequent SP formation.

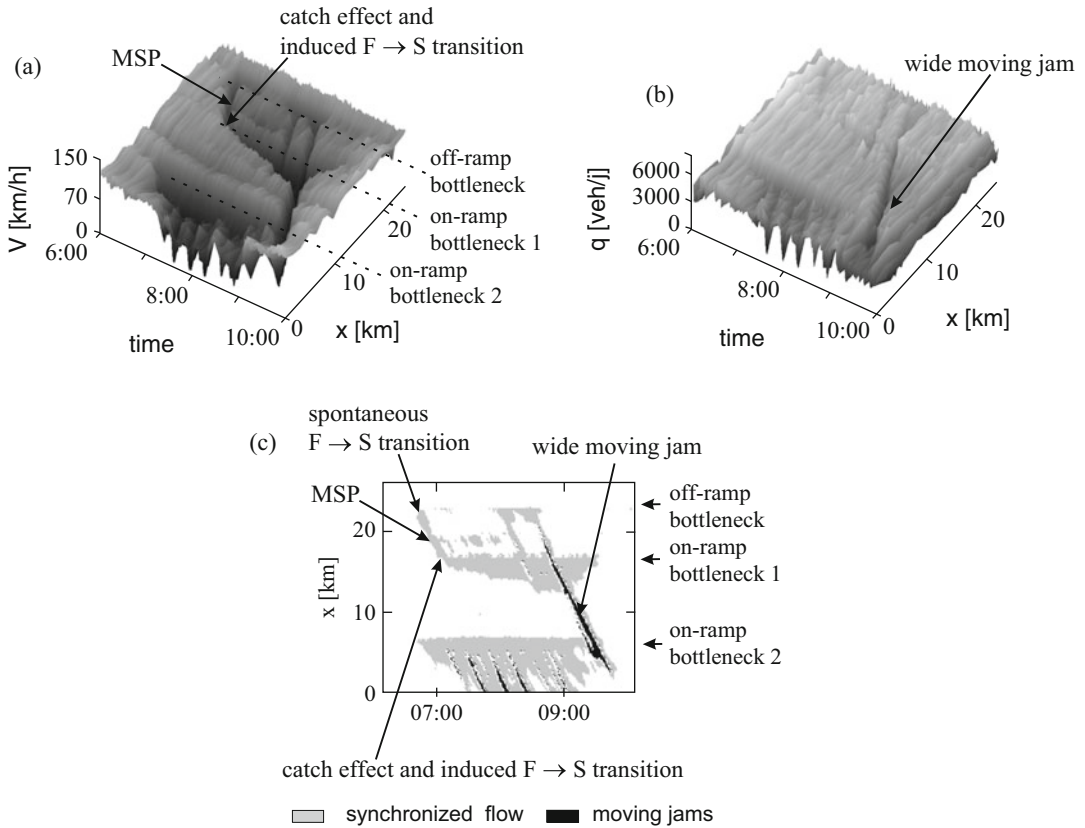
General Congested Patterns

A congested pattern at a hypothetical isolated bottleneck, which occurs due to the $F \rightarrow S \rightarrow J$ transitions, is called a *general pattern* (GP):

- A GP is a congested pattern, which consists of synchronized flow upstream of the bottleneck and wide moving jams that emerge spontaneously in that synchronized flow (Kerner 2004b).

Thus, within the GP, there are two traffic phases of congested traffic, the synchronized flow and wide moving jam traffic phases.

A usual empirical scenario of the emergence of a GP at the bottleneck is as follows. Firstly, due to an $F \rightarrow S$ transition at the bottleneck, synchronized flow occurs at the bottleneck. Then, the pinch



Spatiotemporal Features of Traffic Congestion,

Fig. 44 Empirical example of MSP occurring at off-ramp bottleneck. While propagating upstream, the MSP is caught at an upstream bottleneck (on-ramp bottleneck 1) with subsequent induced $F \rightarrow S$ transition at on-ramp

bottleneck 1: (a, b) Average speed (a) and flow rate (b) in space and time. (c) Graph of (a) with overview of free flow (white), synchronized flow (gray), and moving jams (black). On-ramp bottlenecks 1 and 2 are the same as those in Fig. 1. (Adapted from Kerner 2004b)

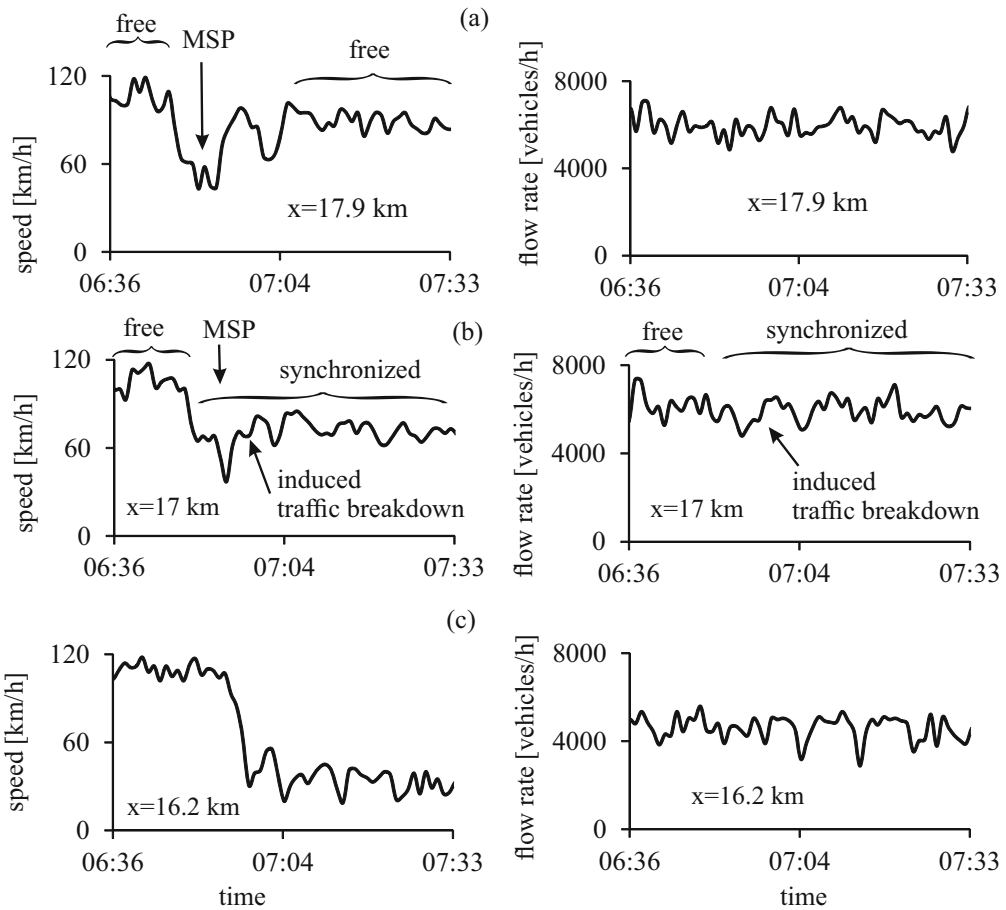
effect occurs in this synchronized flow: due to the $S \rightarrow J$ instability, narrow moving jams emerge spontaneously in the pinch region of synchronized flow (in more details, features of the pinch effect have been considered in section “Moving Jam Emergence in Framework of Three-Phase Traffic Theory”).

The amplitudes of some of the narrow moving jams grow, while the jams propagate upstream within the pinch region. Some of the narrow moving jams (or each of them) transform into wide moving jams at the upstream boundary of the pinch region of synchronized flow. Thus, the width (in the longitudinal direction) of the pinch region of synchronized flow denoted by $L^{(\text{pinch})}$ is determined by a road location at which a narrow

moving jam has transformed into a wide moving jam, i.e., an $S \rightarrow J$ transition has occurred. Because the transformation of different narrow moving jams into wide moving jams can occur at different road locations, the location of the upstream front of the pinch region of synchronized flow can perform complex spatial oscillations over time. Therefore, the width of the pinch region of synchronized flow $L^{(\text{pinch})}$ in a GP can be a complex time-function. The mean width of the pinch region of synchronized flow in a GP denoted by $L_{\text{mean}}^{(\text{pinch})}$ is spatially limited. The value $L_{\text{mean}}^{(\text{pinch})}$ does not depend on traffic demand on the main road upstream of the GP.

Thus, the GP structure consists of the pinch region of synchronized flow (labeled by “pinch

Induced traffic breakdown caused by MSP propagation



Spatiotemporal Features of Traffic Congestion, Fig. 45 Continuation of Fig. 44. Empirical characteristics of induced traffic breakdown caused by the propagation of MSP shown in Fig. 44a, c: (a–c) empirical 1-min average speed (left column) and flow rate (right column) as time-functions measured by detectors installed on freeway

A5-South downstream of the bottleneck ($x = 17.9$ km) (a), at the bottleneck location ($x = 17$ km) (b), and upstream of the bottleneck ($x = 16.2$ km) (c). Data was measured on April 20, 1998. Free flow, synchronized flow, and the MSP are labeled by, respectively, “free,” “synchronized,” and “MSP”

region” in Fig. 47) and a sequence of wide moving jams upstream (labeled by “region of wide moving jam” in Fig. 47).

Types of GPs at Isolated Bottleneck

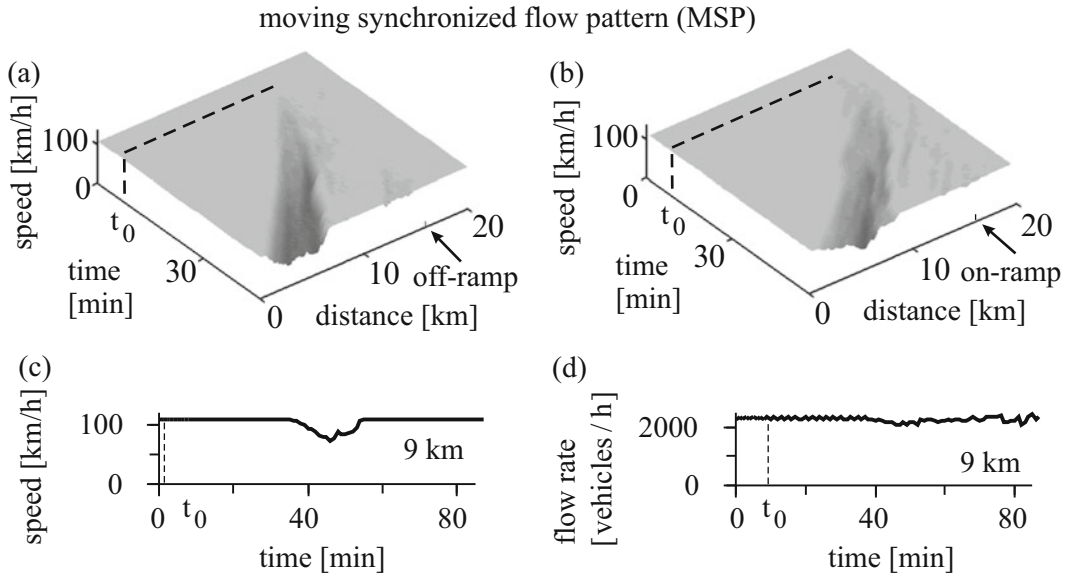
There can be the following types of GPs at an isolated bottleneck:

1. A GP in which the synchronized flow region is upstream bordered by the region of wide moving jams. The upstream front of the GP is related to the upstream front of the farthest

upstream wide moving jam in the region of wide moving jams. This case is symbolically shown in Fig. 47a.

2. A GP in which the upstream front is related to the upstream front of synchronized flow. In this case, the region of wide moving jams is within the synchronized flow region. This case is symbolically shown in Fig. 47b.

Obviously, more complex GPs can be possible that can be considered a combination of these two GP types.



Spatiotemporal Features of Traffic Congestion, Fig. 46 Simulated moving synchronized flow patterns (MSP) at off-ramp (a) and on-ramp (b–d) bottlenecks in the framework of the three-phase theory (Kerner and Klenov 2003): (a, b) Averaged vehicle speed in the MSP

in space and time. (c, d) Speed (c) and flow rate (d) as time-functions at a fixed freeway location. t_0 is a time instant at which the off-ramp outflow in (a) and the on-ramp inflow in (b–d) have been switched on in simulations. $x_{\text{off}} = 16$ km, $x_{\text{on}} = 16$ km. (Adapted from Kerner 2004b)

In the GP of type 1 (Fig. 47a), the upstream front of synchronized flow coincides with the upstream front of the pinch region. This upstream front of synchronized flow is determined by the location where a narrow moving jam has just transformed into a wide moving jam, i.e., where an $S \rightarrow J$ transition occurs. The upstream front separates synchronized flow downstream from the region of wide moving jams upstream. Thus, for the GP of type 1, the width of synchronized flow in the GP denoted by L_{syn} in Fig. 47 coincides with the width of the pinch region of synchronized flow: $L_{\text{syn}} = L^{(\text{pinch})}$.

A successive process of narrow moving jam transformation into wide moving jams at the upstream front of synchronized flow causes a region of wide moving jams in the GP. Due to upstream wide jam propagation, the region of wide moving jams is continuously widening upstream. Consequently, the quantity of wide moving jams within the region of these jams can increase over time. Between wide moving jams, both synchronized flow and free flow can be

formed. These flows will be considered a part of the region of wide moving jams.

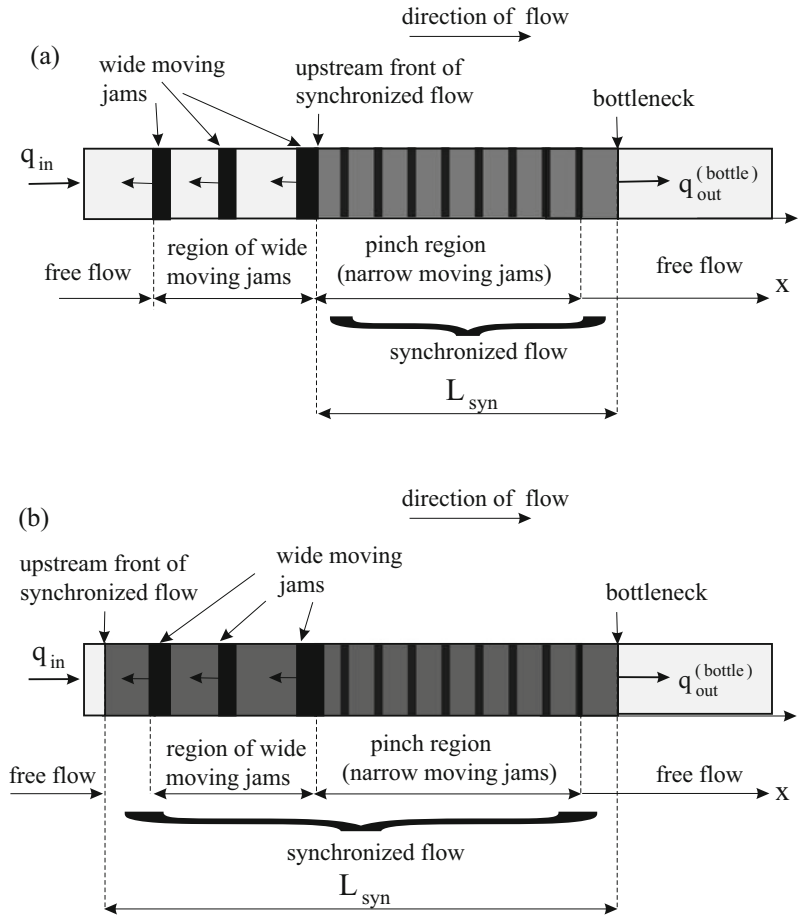
In the GP of type 2 (Fig. 47b), the upstream front of synchronized flow is continuously widening upstream. Therefore, the width of synchronized flow L_{syn} in the GP increases over time rather than being spatially limited. In other words, in contrast to a GP of type 1, the upstream front of synchronized flow in a GP of type 2 is not determined by the location of an $S \rightarrow J$ transition (wide moving jam emergence). Firstly, synchronized flow propagates continuously upstream of the bottleneck, and only later wide moving jams emerge in that synchronized flow.

Some Empirical Scenarios of GP Formation

It must be noted that over time the average speed in synchronized flow of a WSP can decrease, and wide moving jams can begin to emerge spontaneously in synchronized flow. In other words, the WSP can transform into a GP (Fig. 48). In accordance with GP definition, the empirical GP in Fig. 48a, b is a congested pattern at an isolated

Spatiotemporal Features of Traffic Congestion,

Fig. 47 Symbolical spatial structures of GPs at an isolated effectual bottleneck at a given time instant. (a) GP of type 1. (b) GP of type 2. (Adapted from Kerner 2004b)



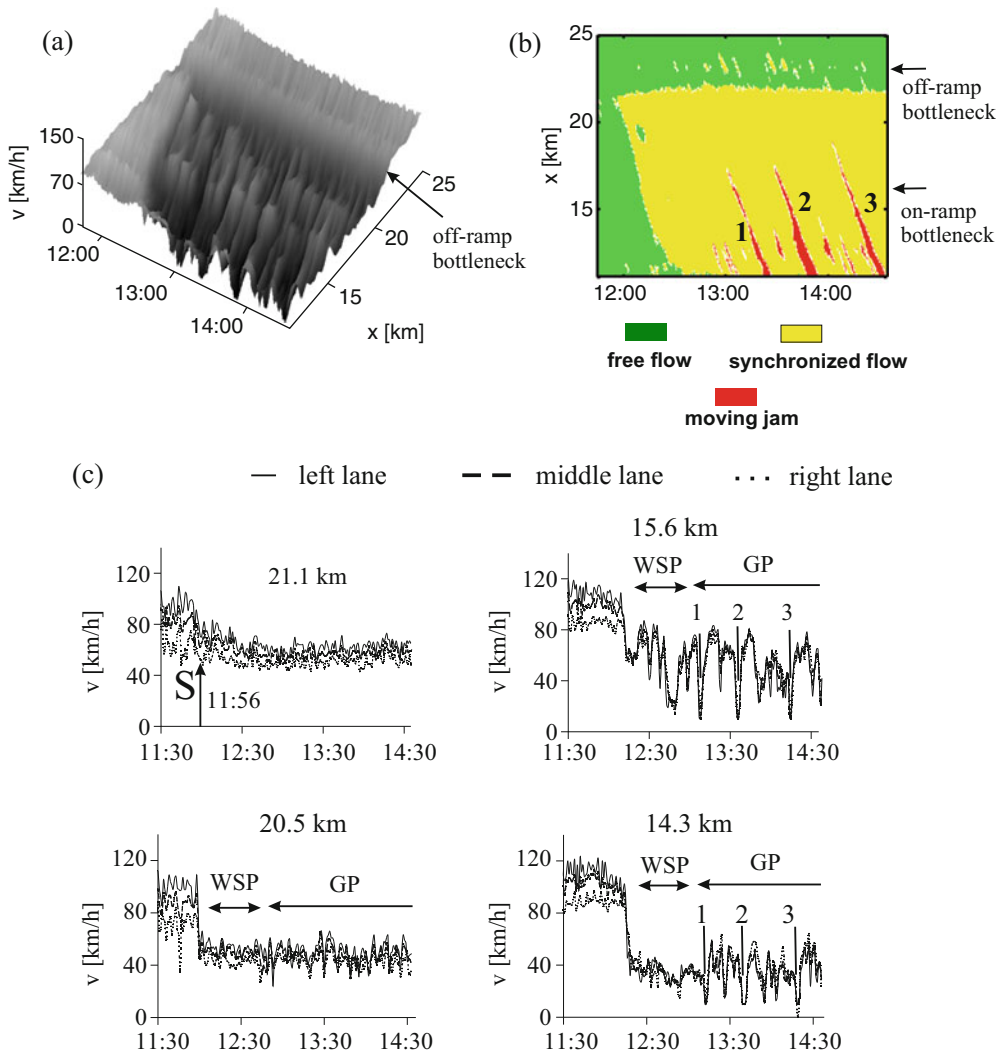
bottleneck, which consists of synchronized flow whose downstream front is fixed at the bottleneck and wide moving jams that emerge spontaneously in this synchronized flow upstream of the bottleneck.

In Fig. 48, WSP transformation into the GP occurs downstream of an upstream on-ramp bottleneck (labeled by “on-ramp bottleneck” in Fig. 48b). For this reason, we can qualitatively consider the congested pattern that is upstream of the downstream off-ramp bottleneck (labeled by “off-ramp bottleneck” in Fig. 48a, b and downstream of the upstream on-ramp bottleneck as an empirical example of a GP at the off-ramp bottleneck. Because the moving jams labeled 1, 2, and 3 propagate through the upstream bottleneck while maintaining their mean velocities of the jam downstream front (Fig. 48b), in accordance

with the phase definition [J], these moving jams are indeed wide moving jams.

However, we must note that in many cases, upstream congestion (i.e., synchronized flow and wide moving jams in Fig. 48a, b upstream of the on-ramp bottleneck) can influence downstream congested pattern parameters considerably (section “Complex Congested Patterns and Pattern Interaction at Adjacent Bottlenecks”). For this reason, the above consideration of the empirical congested pattern between the adjacent off-ramp and on-ramp bottlenecks (Fig. 48) as a GP can only show some qualitative features of GPs at an isolated bottleneck.

Another empirical example of a GP, which is formed at an on-ramp bottleneck before wide moving jams or/and synchronized flow of this GP reach an eventual upstream bottleneck on the



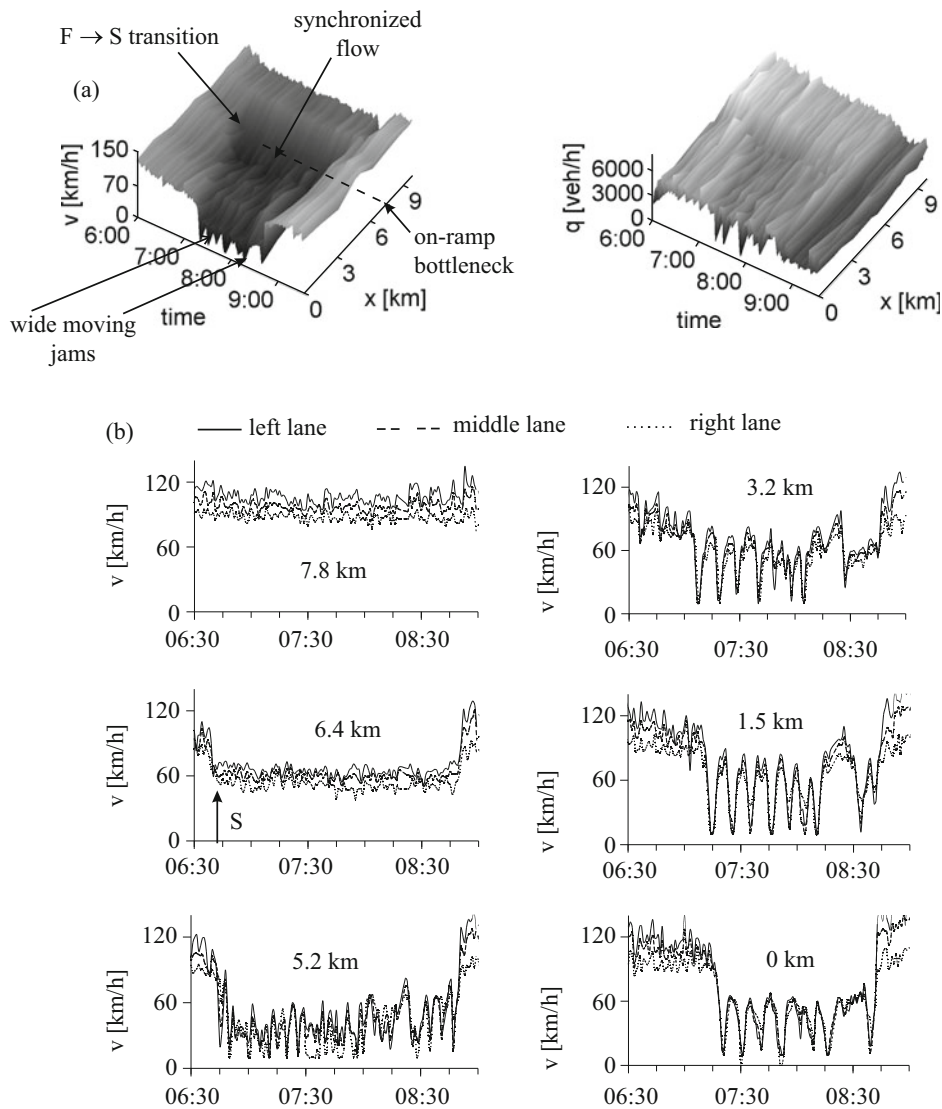
Spatiotemporal Features of Traffic Congestion, Fig. 48 Empirical example of WSP transformation into an GP over time: (a) Average speed in space and time. (b) Graph of (a) with overview of free flow (green),

synchronized flow (yellow) and moving jams labeled 1, 2, and 3 (red). (c) Time dependence of speed at some fixed freeway locations; arrows S show the time instant of traffic breakdown. (Adapted from Kerner 2004b)

freeway, is shown in Fig. 49. We see that firstly, synchronized flow is formed upstream of the bottleneck (labeled by arrows $F \rightarrow S$ and S). Later, and at other freeway locations, wide moving jams emerge in that synchronized flow. These wide moving jams propagate further upstream while maintaining the velocity of their downstream fronts. Free flow remains downstream of the bottleneck (location 7.8 km).

We can see that qualitative features of this GP are the same as those in Fig. 48. In both cases, firstly, an

$F \rightarrow S$ transition occurs at the bottleneck (arrow S in Figs. 48c and 49b). Synchronized flow propagates upstream. Later, the pinch effect is realized: narrow moving jams emerge spontaneously in the synchronized flow upstream of the location of the initial traffic breakdown. These narrow moving jams propagate upstream growing in amplitude. Some of the jams transform into wide moving jams, i.e., $S \rightarrow J$ transitions occur. In the example, the mean time interval between the emergence of a narrow moving jam and the jam transformation into a wide moving



Spatiotemporal Features of Traffic Congestion, Fig. 49 Empirical example of GP at on-ramp bottleneck. (a) Average vehicle speed (left) and total flow rate (right)

across the freeway in space and time. (b) Speed in the three freeway lanes at different locations (1 min average traffic data). (Adapted from Kerner 2004b)

one is about 10 min, i.e., about ten times longer than the mean duration of F $\rightarrow$ S transitions, which are about 1 min in duration.

However, there are also some differences in GP formation in Figs. 48 and 49. In Fig. 48, WSP exists during a long enough time interval before the pinch effect occurs within the WSP leading to WSP transformation into the GP. In contrast, in Fig. 49, the pinch effect occurs in synchronized flow after a short time interval. Secondly, in

Fig. 48, the frequency of wide moving jam emergence in synchronized flow is considerably smaller than the one for the GP shown in Fig. 49.

The differences in GP formation that have been mentioned above result from different pinch effect features in these two cases (Figs. 48 and 49). It has been found that the stronger the self-compression of synchronized flow in the pinch region is (i.e., the more the increase in the density of the synchronized flow), the more likely narrow moving

jams emerge, and the larger the frequency of this moving jam emergence is. This is in accordance with pinch effect features discussed in section “Moving Jam Emergence in Framework of Three-Phase Traffic Theory.” In the first case of the off-ramp bottleneck shown in Fig. 48, the pinch effect is weak (the so-called “weak congestion” case, see section “Diagram of Congested Patterns at Isolated Bottlenecks”). In this case, a relatively small increase in the density of the synchronized flow within the initial WSP occurs. As a result, the frequency of moving jam emergence is small; the frequency depends on the percentage of vehicle leaving the main road to the off-ramp.

In contrast, in Fig. 49, the pinch effect is relatively strong, i.e., the vehicle density increases considerably in the pinch region of synchronized flow; the frequency of moving jam emergence is larger than in the former case. It has been found that if the flow rates upstream of the on-ramp bottleneck increase, then a case of the so-called strong congestion within the pinch region can be reached (see section “Diagram of Congested

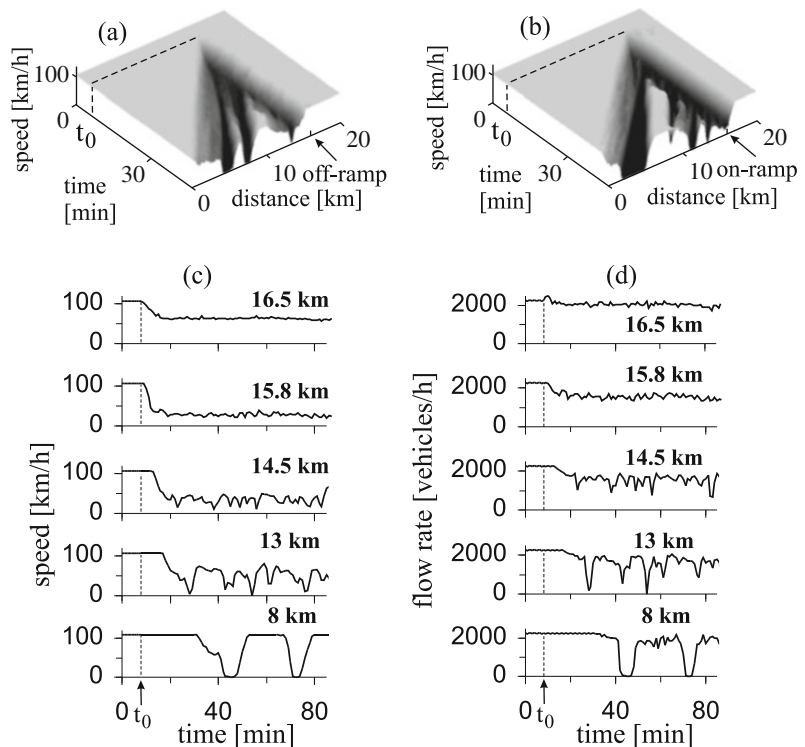
Patterns at Isolated Bottlenecks”). In this case, regardless of a further increase in the on-ramp flow rate, the flow rate within the pinch region on the main road reaches a limit (minimum) flow rate, and the frequency of moving jam emergence reaches the maximum value. This is valid, if bottleneck characteristics and traffic parameters like weather, percentage of long vehicles, and other road conditions do not change considerably.

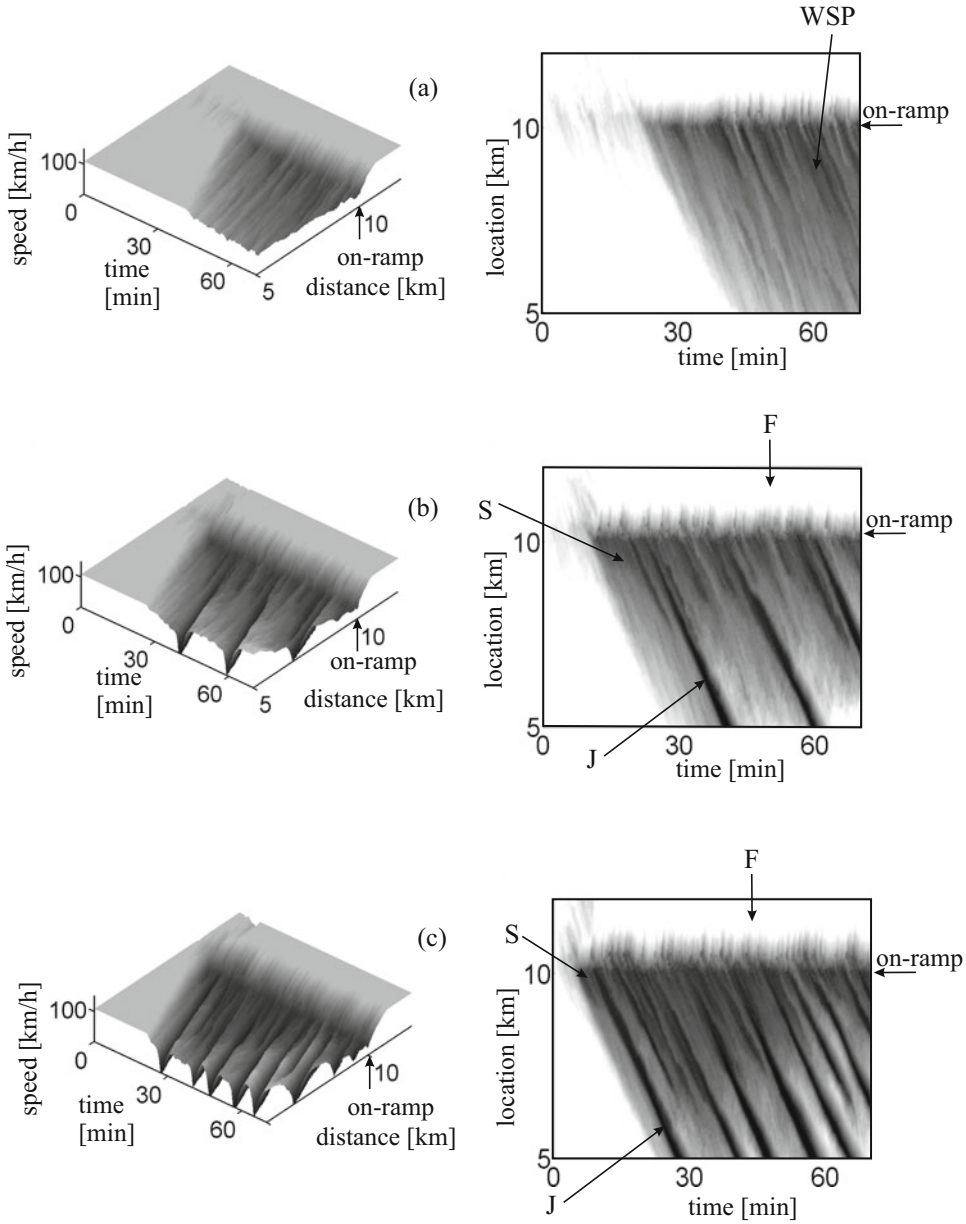
These and other empirical spatiotemporal features of GP formation have been explained in the three-phase theory (Fig. 50). A more detailed consideration of empirical and theoretical spatiotemporal features of GPs has been done in the book Kerner (2004b). In particular, readers can find in Kerner (2004b) the following features of GPs that have not been mentioned above:

- (i) There are GPs whose upstream front is associated with the upstream front of the farthest upstream wide moving jam.
- (ii) In accordance with section “Hypotheses About Nucleation Nature of Moving Jam

Spatiotemporal Features of Traffic Congestion,

Fig. 50 Simulations of GP at off-ramp (a) and on-ramp bottlenecks (b–d). (a, b) Average vehicle speed in space and time. (c, d) Speed (c) and flow rate (d) at different locations (1 min data averaging) for the GP in (b). t_0 is a time instant at which the off-ramp outflow (a) and the on-ramp inflow (b–d) have been switched on in simulations. (Adapted from Kerner 2004b)





Spatiotemporal Features of Traffic Congestion, Fig. 51 Simulations of the occurrence of heavy traffic congestion at on-ramp bottleneck on single-lane road. Speed in space and time (left column) and the same data presented by regions with variable shades of gray (shades of gray vary from white to black when the speed decreases from 105 km/h (white) to zero (black)) (right column). In simulations, the same flow rate in free flow upstream of the on-ramp $q_{in} = 2000$ vehicles/h and different on-ramp

inflow rates have been used: (a) $q_{on} = 250$, (b) $q_{on} = 350$, and (c) $q_{on} = 550$ vehicles/h. F, free flow; S, synchronized flow; J, wide moving jam; WSP, widening synchronized flow pattern. $x_{on} = 10$ km and $x_{on}^{(e)} = 10.3$ km are, respectively, the beginning and the end of the merging region of the on-ramp within which vehicles can merge from the on-ramp onto the main road. (Adapted from Kerner 2017a)

Emergence in Synchronized Flow,” points related to the pinch region of a GP lie above the line J in the flow–density plane, i.e., these points are associated with metastable synchronized flow states with respect to moving jam emergence. This explains narrow moving jam emergence in the pinch region.

- (iii) After a narrow moving jam has transformed into a wide moving jam, this wide moving jam can suppress the growth of the downstream narrow moving jam (Fig. 50c, $x = 13$ km). This jam suppression effect occurs only if the narrow moving jam is close to the downstream wide moving jam front.
- (iv) Distances between different narrow moving jams in the pinch region can be very different to one another.

It must be noted that the well-known and very old terms “stop-and-go traffic” and “freeway traffic oscillations,” which are related to a sequence of moving traffic jams, will not be used in this article. For a traffic observer, both a sequence of narrow moving jams and a sequence of wide moving jams constitute either “stop-and-go traffic” or “freeway traffic oscillations.” However, as already mentioned, narrow moving jams belong to the synchronized flow phase, whereas wide moving jams belong to the qualitative different wide moving jam phase.

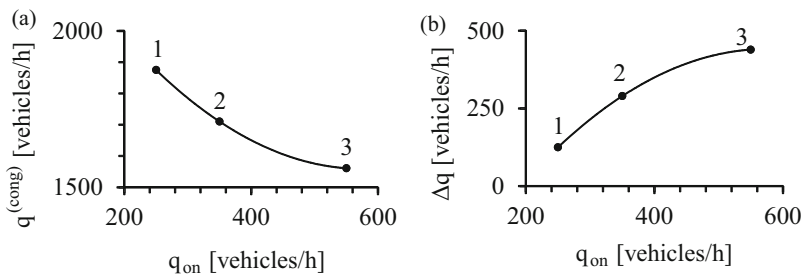
Heavy Traffic Congestion

To understand the term “heavy traffic congestion,” we consider simulations of traffic breakdown ($F \rightarrow S$ transition) with the resulting formation of congested patterns on a single-lane road with an on-ramp bottleneck (Figs. 51 and 52). The flow rate in free flow at the bottleneck is denoted by q_{sum} , where $q_{\text{sum}} = q_{\text{in}} + q_{\text{on}}$, q_{in} is the flow rate in free flow upstream of the on-ramp bottleneck, and q_{on} is the on-ramp inflow rate. The flow rate q_{sum} that in simulations (Figs. 51 and 52) is equal to the flow rate in free flow measured before traffic breakdown has occurred at the bottleneck is often called the predischARGE flow rate. We consider the flow rate drop

$$\Delta q = q_{\text{sum}} - q_{\text{out}}^{(\text{bottle})} \quad (7)$$

that occurs at the bottleneck due to traffic breakdown with the resulting congested pattern formation. In (7), $q_{\text{out}}^{(\text{bottle})}$ is the flow rate in free flow downstream of the bottleneck at which congested traffic pattern is realized; this flow rate is called the discharge flow rate. Respectively, we denoted by $q^{(\text{cong})}$ the mean flow rate within congested traffic measured at a short distance (about 0.3–1 km) upstream of the bottleneck.

When the flow rate q_{in} is large enough but the on-ramp inflow rate q_{on} is not large, the WSP occurs at the bottleneck (Fig. 51a). Both empirical results and simulations made in the framework of



Spatiotemporal Features of Traffic Congestion, Fig. 52 Flow–flow characteristics for traffic patterns shown in Fig. 51: (a) Dependence of the mean flow rate in congested traffic $q^{(\text{cong})}$ on the on-ramp inflow rate q_{on} . (b) Dependence of the mean flow rate drop due to congested traffic Δq (7) in free flow downstream of the bottleneck (at $x = 11$ km) on the on-ramp inflow rate q_{on} .

The mean flow rate in congested traffic $q^{(\text{cong})}$ averaged over 60 min is measured by a virtual detector at the road location $x = 9.5$ km that is 500 m upstream of the beginning of the merging region of the on-ramp $x_{\text{on}} = 10$ km. Point 1 is related to WSP in Fig. 51a, point 2 is related to GP in Fig. 51b, and point 3 is related to GP in Fig. 51c. (Adapted from Kerner 2017a)

the three-phase theory show that the flow rate $q^{(\text{cong})}$ within the WSP is usually not considerably smaller than the flow rate q_{in} . Respectively, in the case of the WSP shown in Fig. 51a, we have found that the flow rate $q^{(\text{cong})}$ is only 125 vehicles/h smaller than the flow rate $q_{\text{in}} = 2000$ vehicles/h (Fig. 52a). For this reason, in the case of WSP formation (Fig. 51a), the flow rate drop Δq (7) is about 5.6% of the flow rate q_{sum} (Fig. 52b).

When at the same flow rate q_{in} the on-ramp inflow rate q_{on} increases, the average speed in synchronized flow decreases. In accordance with results of sections “Moving Jam Emergence in Framework of Three-Phase Traffic Theory” and “General Congested Patterns,” in synchronized flow of a smaller speed, the pinch effect is realized (Fig. 51b, c): The vehicle density increases, and narrow moving jams ($S \rightarrow J$ instability) appear randomly in the pinch region of synchronized flow. Some of the narrow moving jams grow continuously over time. Such growing narrow moving jams are nuclei for $S \rightarrow J$ transitions (Kerner 2004b). Finally, the growing narrow moving jams transform into wide moving jams resulting in a GP that consists of the synchronized flow and wide moving jam phases.

The mean flow rate $q^{(\text{cong})}$ is measured within the pinch region of synchronized flow of the GP. In a theory of GPs (Kerner 2004b), this flow rate has been denoted by $q^{(\text{pinch})}$ (see section “Diagram of Congested Patterns at Isolated Bottlenecks” below). In this section, we use a general designation $q^{(\text{cong})}$ used in Kerner (2008) for the mean flow rate in congested traffic.

The larger the on-ramp inflow rate q_{on} at a given flow rate q_{in} , the smaller the mean flow rate in congested traffic $q^{(\text{cong})}$. The value of $q^{(\text{cong})}$ can become considerably smaller than the flow rate q_{in} (points 2 and 3 in Fig. 52a). Correspondingly, the flow rate drop $\Delta q = q_{\text{sum}} - q_{\text{out}}^{(\text{bottle})}$ (7) between the predischage flow rate q_{sum} and the discharge flow rate $q_{\text{out}}^{(\text{bottle})}$ increases considerably (points 2 and 3 in Fig. 52b):

- Traffic congestion at a bottleneck that results in a large value of the flow rate drop Δq (7) at the bottleneck can be considered *heavy traffic congestion*.

We see that there is no distinct criterion in the above definition of heavy traffic congestion. The sense of this definition is to explain below the fundamental requirement for the reliability of intelligent transportation systems (ITS).

The Fundamental Requirement for Reliability of Intelligent Transportation Systems (ITS)

The objective of this section is to show that the fundamental requirement for the reliability of ITS (e.g., automated and cooperative driving vehicles, methods and strategies for traffic control, dynamic traffic assignment, as well as dynamic network optimization) is the consistence of ITS with the empirical nucleation nature of traffic breakdown at a road bottleneck.

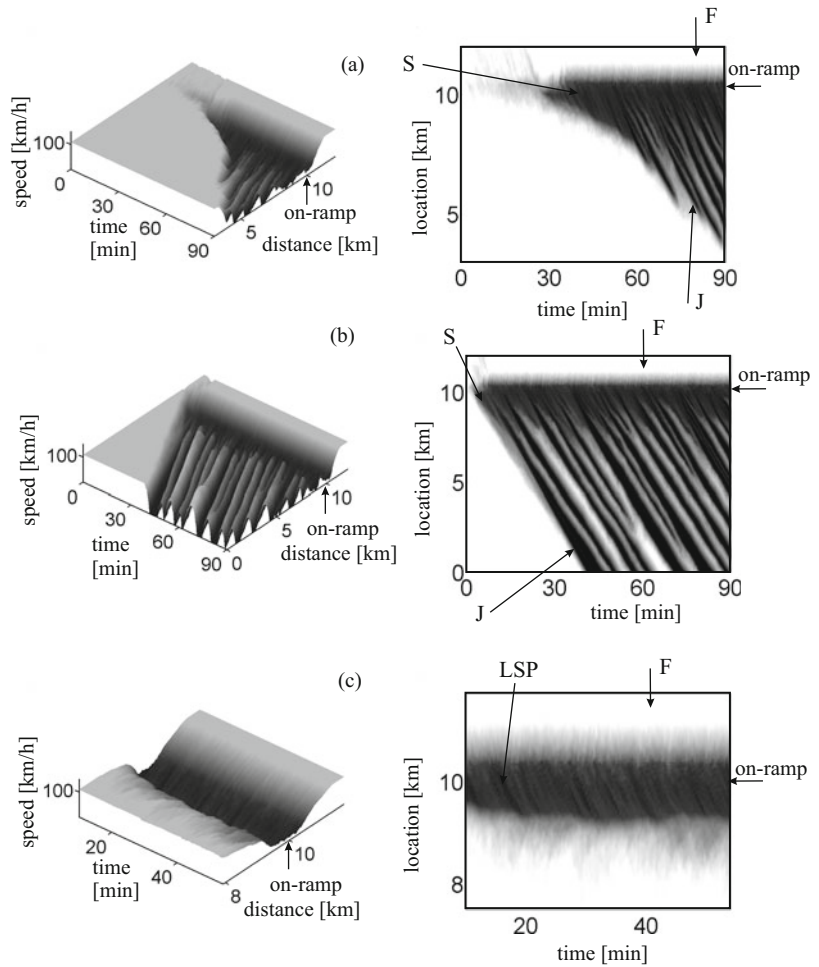
Indeed, after heavy traffic congestion has already developed upstream of a network bottleneck within a congested traffic pattern resulting from traffic breakdown, it is very difficult with the use of ITS to dissolve the congested traffic pattern. This leads to a well-known conclusion that the basic objective of any reliable ITS is to prevent traffic breakdown in a traffic or transportation network. However, to prevent real traffic breakdown, ITS must be consistent with the nucleation nature of traffic breakdown at the bottleneck. However, the classical traffic theories and models are not consistent with the nucleation nature of traffic breakdown at the bottleneck. This is the reason why the author states that the classical traffic theories and models cannot be applied for reliable traffic control, dynamic traffic assignment, dynamic traffic optimization, and other ITS applications (more detailed explanations of the criticism of the classical traffic theories and models can be found in the book (Kerner 2017a) as well as in the reviews (Kerner 2017b, 2018a) of this Encyclopedia).

In general, the larger the on-ramp inflow rate q_{on} , the smaller the mean flow rate in congested traffic $q^{(\text{cong})}$ (Fig. 52a). Therefore, the larger the on-ramp inflow rate q_{on} , the more difficult to dissolve traffic congestion through the reduction of the flow rate q_{in} in free flow upstream of the on-ramp bottleneck. A numerical proof of this statement is presented in Figs. 53 and 54.

In Fig. 54, the dependence of the flow rate in free flow downstream of the on-ramp bottleneck

Spatiotemporal Features of Traffic Congestion,

Fig. 53 Simulations of congested patterns at on-ramp bottleneck on single-lane road related to a given on-ramp inflow rate q_{on} and a variable value of the flow rate q_{in} in free flow upstream of the congested patterns. Speed in space and time (left column) and the same data presented by regions with variable shades of gray (shades of gray vary from white to black when the speed decreases from 105 km/h (white) to zero (black)) (right column). Simulations have been made at the same on-ramp inflow rate $q_{on} = 770$ vehicles/h and different flow rates q_{in} in free flow upstream of congested patterns: (a) GP at $q_{in} = 1500$ vehicles/h. (b) GP at $q_{in} = 2000$ vehicles/h. (c) LSP at $q_{in} = 1200$ vehicles/h. (Adapted from Kerner 2017a)



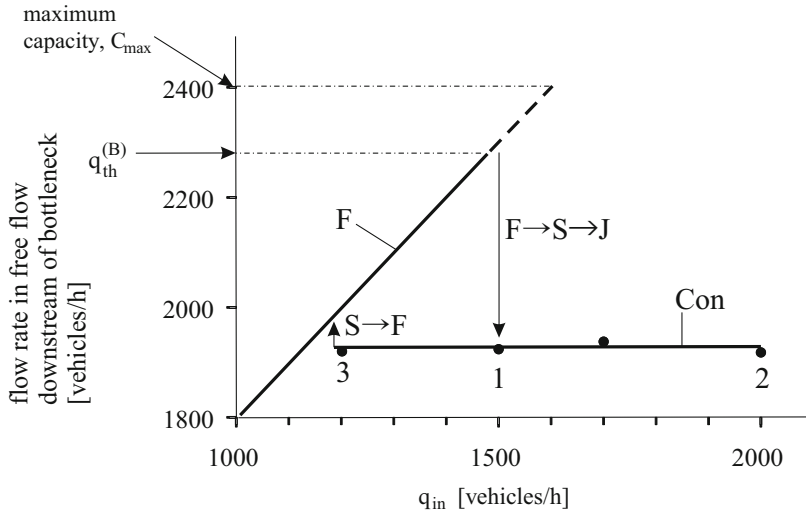
on the flow rate q_{in} in free flow upstream of the bottleneck calculated at a given on-ramp inflow rate q_{on} for congested patterns shown in Fig. 53 is presented.

Curve F in Fig. 54 is related to free flow at the on-ramp bottleneck. The flow rate in free flow downstream of the on-ramp bottleneck is equal to $q_{sum} = q_{in} + q_{on}$. A dashed part of curve F in Fig. 54 is related to metastable free flow with respect to F→S transition (traffic breakdown).

In the case under consideration (Figs. 53 and 54), a GP emerges spontaneously due to a sequence of F→S→J transitions (arrow F→S→J in Fig. 54). It should be noted that all wide moving jams within the GP shown in Fig. 53a dissolve during their upstream propagation. This is because the flow rate $q_{in} =$

1500 vehicles/h is smaller than the flow rate in the jam outflow q_{out} that under chosen model parameter is equal to $q_{out} = 1808$ vehicles/h. Nevertheless, the upstream front of the GP propagates upstream over time (Fig. 53a).

A solid curve labeled by “Con” in Fig. 54 is related to the dependence of the discharge flow rate $q_{out}^{(bottle)}$ from congested patterns at the bottleneck on the flow rate q_{in} in free flow upstream of the patterns. We can see that the discharge flow rate $q_{out}^{(bottle)}$ (curve “Con” in Fig. 54) is nearly some constant value that is independent of the flow rate q_{in} and on pattern characteristics. This result can be understood, if we consider the dependence of the mean flow rate within congested patterns $q^{(cong)}$ on the flow rate q_{in} (Fig. 55): the mean



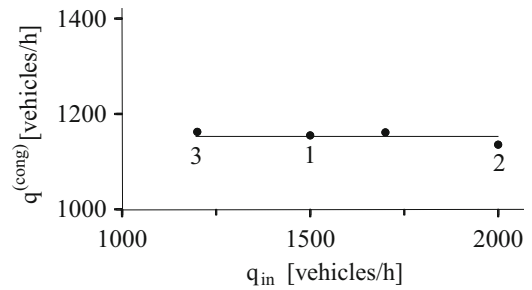
Spatiotemporal Features of Traffic Congestion, Fig. 54 Continuation of Fig. 53. Simulated flow–flow characteristic for the emergence and evolution of congested patterns at on-ramp bottleneck shown in Fig. 53. Curve F is related to free flow. Curve labeled by “Con” is related to the discharge flow rate $q_{out}^{(bottle)}$ from congested patterns measured through the use of a virtual detector at location $x = 11$ km (1 km downstream of the

beginning of the on-ramp merging region); averaging time interval of the flow rate $q_{out}^{(bottle)}$ is equal to 60 min. Point 1 on curve “Con” is related to GP shown in Fig. 53a. Point 2 on curve “Con” is related to GP shown in Fig. 53b. Point 3 on curve “Con” is related to LSP shown in Fig. 53c. $q_{on} = 770$ vehicles/h. Calculated values $q_{th}^{(B)} = 2280$ vehicles/h, $C_{max} = 2400$ vehicles/h, $q_{out}^{(bottle)} \approx 1930$ vehicles/h. (Adapted from Kerner 2017a)

flow rate within congested patterns $q^{(cong)}$ is nearly some constant value that is independent on the flow rate q_{in} .

If the flow rate q_{in} increases to 2000 vehicles/h, neither the mean flow rate $q^{(cong)}$ within the pinch region of the GP nor the discharge flow rate $q_{out}^{(bottle)}$ have changed (point 2 on curve “Con” in Fig. 54). We observe only that the spatiotemporal structure of the GP becomes more complex (Fig. 53b). Thus, through congestion occurrence the flow rate upstream of the bottleneck can reduce considerably from $q_{in} = 2000$ vehicles/h in free flow to $q^{(cong)} \approx 1160$ vehicles/h in congested traffic (point 2 in Fig. 55). Therefore, we can call this traffic congestion as heavy traffic congestion.

If the flow rate q_{in} decreases to 1200 vehicles/h, no wide moving jams emerge within synchronized flow any more: the GP transforms into an SP. However, also in this case, neither the mean flow rate $q^{(cong)}$ within the SP nor the discharge flow rate $q_{out}^{(bottle)}$ has changed (point 3 on curve “Con” in Fig. 54). Because the flow rate $q_{in} = 1200$ vehicles/h is only about 3% larger than $q^{(cong)} \approx$



Spatiotemporal Features of Traffic Congestion, Fig. 55 Continuation of Fig. 53. Dependence of the flow rate within congested traffic $q^{(cong)}$ on the flow rate in free flow q_{in} upstream of congested patterns shown in Fig. 53: point 1 is related to GP in Fig. 53a, point 2 is related to GP in Fig. 53b, point 3 is related to LSP in Fig. 53c. Calculated value $q^{(cong)} \approx 1160$ vehicles/h. (Adapted from Kerner 2017a)

1160 vehicles/h, the upstream front of the SP propagates upstream extremely slow: the width (in the longitudinal direction) of the SP does not almost change over time. Therefore, we can consider this SP as an LSP.

Thus, heavy traffic congestion cannot be dissolved even when the flow rate q_{in} in free flow upstream of the bottleneck decreases from 2000 vehicles/h to 1200 vehicles/h. When the flow rate q_{in} becomes smaller than the mean flow rate within the pinch region of synchronized flow $q^{(cong)} \approx 1160$ vehicles/h, an S→F transition occurs at the bottleneck (arrow labeled by S→F in Fig. 54). At $q_{in} < q^{(cong)}$, no traffic congestion can occur at the bottleneck.

The above analysis of traffic congestion leads to the conclusion that rather than controlling traffic congestion in the network, ITS should either prevent traffic breakdown or limit the development of traffic congestion in the network. To prevent traffic breakdown, ITS in the framework of the three-phase theory are needed. This is because none of the classical traffic theories can explain the empirical nucleation nature of real traffic breakdown at highway bottlenecks:

- The fundamental requirement for the reliability of intelligent transportation systems (ITS) is that ITS should be consistent with the empirical nucleation nature of traffic breakdown at a highway bottleneck.

Diagram of Congested Patterns at Isolated Bottlenecks

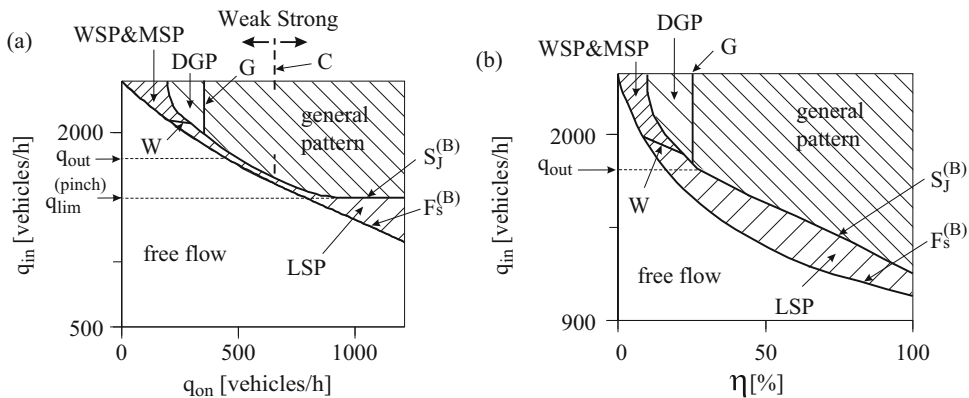
A diagram of different congested patterns at an on-ramp bottleneck gives regions of congested

pattern occurrence in the flow–flow plane whose coordinates are the flow rate in free flow on the main road upstream of the bottleneck q_{in} and the flow rate to the on-ramp q_{on} (Fig. 56a) (Kerner and Klenov 2002, 2003; Kerner et al. 2002). There are two main boundaries in this diagram, $F_S^{(B)}$ and $S_J^{(B)}$.

At any point (q_{on}, q_{in}) related to the boundary $F_S^{(B)}$ in the diagram, traffic breakdown (F→S transition) occurs at the bottleneck with probability $P^{(B)} = 1$ during a given time interval T_{ob} for observing free flow at the bottleneck. We note that the maximum highway capacity of free flow at the bottleneck C_{max} is equal to the flow rate at the bottleneck at which traffic breakdown occurs at the bottleneck with probability $P^{(B)} = 1$ during the time interval T_{ob} (see Encyclopedia article Kerner 2018a). This means that at the boundary $F_S^{(B)}$ highway capacity is equal to the maximum highway capacity C_{max} .

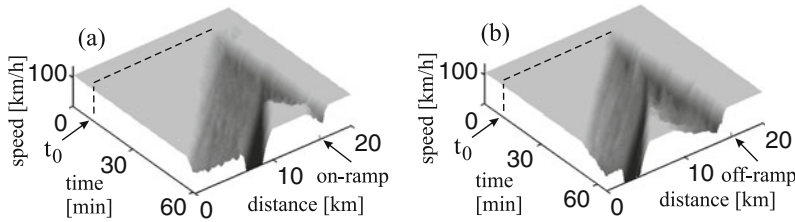
It should be noted that in Fig. 56 we show some simplified diagrams of congested patterns. Here we do not show a threshold diagram boundary $F_{th}^{(B)}$ that is left of the boundary $F_S^{(B)}$. Between the boundaries $F_S^{(B)}$ and $F_{th}^{(B)}$ free flow is in a metastable state with respect to traffic breakdown (see explanations of congested pattern features for this more general case in the book Kerner 2004b).

Between the boundaries $F_S^{(B)}$ and $S_J^{(B)}$, different SPs (LSP, WSP, and MSP) emerge upstream of the



Spatiotemporal Features of Traffic Congestion, Fig. 56 Diagrams of congested patterns at on-ramp (a) and off-ramp bottlenecks (b). Results of numerical

simulations. SP, synchronized flow pattern; WSP, widening SP; MSP, moving SP; LSP, localized SP; GP, general pattern; DGP, dissolving GP (Adapted from Kerner 2004b)



Spatiotemporal Features of Traffic Congestion, Fig. 57 Dissolving GP (DGP) at on-ramp (a) and off-ramp (b) bottlenecks. Results of numerical simulations.

At $t = t_0$ the on-ramp inflow (a) or the off-ramp outflow (b) has been switched on. (Adapted from Kerner 2004b)

on-ramp (Figs. 41, 43, and 46). Right of the boundary $S_j^{(B)}$, wide moving jams occur spontaneously in synchronized flow, i.e., different GPs appear (Figs. 50 and 57). At the boundary $S_j^{(B)}$ within a given time interval for observing synchronized flow, an S→J transition occurs in synchronized flow that leads to wide moving jam emergence.

Considering regions of SP emergence in the diagram, the following conclusions can be made. The WSP occurs above the boundary W in the diagram in Fig. 56a. Below the boundary W , an LSP occurs. At larger flow rates q_{in} and smaller flow rates q_{on} , an MSP can occur rather than the WSP. The MSP occurs spontaneously on the main road at the on-ramp bottleneck because at a low flow rate q_{on} , the downstream front of synchronized flow can depart from the on-ramp bottleneck. The MSP begins to propagate on the main road as an independent localized structure (Fig. 46) on a homogeneous road. A sequence of MSPs can also occur on the main road at the on-ramp bottleneck.

When right of the boundary $S_j^{(B)}$ at a given q_{in} the flow rate q_{on} increases, the average density within the pinch region of an GP increases too. There is a boundary labeled C in the diagram, which separates weak and strong congestion conditions within the pinch region of GPs. Under weak congestion conditions (left of boundary C in Fig. 56a), the flow rate within the pinch region $q^{(pinch)}$ and the frequency of moving jam emergence in the pinch region depend on the on-ramp inflow rate q_{on} . Contrarily, under strong congestion conditions (right of boundary C in Fig. 56a), the flow rate $q^{(pinch)}$ and the frequency of moving jam emergence reach their limit minimum values that do not depend on the on-ramp

inflow rate q_{on} any more. We denote the limit minimum value of the flow rate within the pinch region by $q_{lim}^{(pinch)}$. We can see that in contrast with the term *heavy traffic congestion* used above for the explanation of the failure of ITS applications based on the classical traffic flow theories, the terms *weak congestion* and *strong congestion* satisfy distinct criteria.

There is a region in the diagram that is between the boundary $S_j^{(B)}$ and the line G within which a dissolving GP (DGP) occurs (Figs. 50a and 57a). In a DGP, after a wide moving jam has been formed, this wide moving jam suppresses the pinch region in the initial GP. The suppression effect can occur when the initial flow rate q_{in} is high enough and the flow rate to the on-ramp q_{on} is lower than some characteristic value that determines the position of the boundary G in the diagram of congested patterns in Fig. 56a.

To understand the physics of the DGP, let us consider flow rates q_{in} that satisfy the condition $q_{in} > q_{out}$ (q_{out} is the flow rate in the wide moving jam outflow under condition that in this jam outflow free flow is formed). After the first wide moving jam has been formed in the GP, the flow rate in the jam outflow cannot be larger than q_{out} . This means that rather than an initial high flow rate q_{in} , a smaller flow rate related to the jam outflow determines the inflow into the pinch region of the GP. At a low enough flow rate to the on-ramp q_{on} (left of the boundary G in Fig. 56a), the pinch region cannot exist at this lower inflow into the pinch region: the pinch region dissolves. As a result of this dissolution effect, either free flow or an LSP appears at the on-ramp (Fig. 57).

In contrast to the on-ramp bottleneck, at an off-ramp bottleneck (Fig. 50a), a GP *only* under weak congestion can be formed. This means that the flow rate $q^{(\text{pinch})}$ and the frequency of moving jam emergence in the pinch region depend on the percentage of vehicles η that want to leave the main road to the off-ramp in the flow rate q_{in} . This is true for all values $\eta < 100\%$.

The pattern diagram at the off-ramp bottleneck, i.e., the regions of pattern formation in the plane (η, q_{in}) (Fig. 56b), qualitatively resembles the diagram at the on-ramp bottleneck (Fig. 56a). However, there is no saturation of the boundary $S_j^{(B)}$ at a higher η . This is associated with the mentioned fact that in GPs only the weak congestion condition occurs at all $\eta < 100\%$. Indeed, all pinch region characteristics in GPs do depend on η .

Note that GP parameters depend on the average flow rate $q^{(\text{cong})}$ within a GP at a bottleneck. The time interval of flow rate averaging of $q^{(\text{cong})}$ is assumed to be considerably longer than time distances between any moving jams within the GP. For the GP, the average flow rate $q^{(\text{cong})}$ is equal to the average flow rate within the pinch region of synchronized flow: $q^{(\text{cong})} = q^{(\text{pinch})}$. In the theory of GPs and the pinch effect (Kerner and Klenov 2002, 2003; Kerner et al. 2002), it was found that the larger the bottleneck influence (called the bottleneck strength) on traffic is, the smaller $q^{(\text{pinch})}$ is. In turn, the smaller $q^{(\text{pinch})}$ is, the larger the mean frequency of wide moving jams in a GP, and the lower the average speed between the wide moving jams.

In particular, for an on-ramp bottleneck at a given flow rate q_{in} , the larger the flow rate to the on-ramp q_{on} is, the larger the bottleneck strength; however, this is only true as long as weak congestion conditions within the pinch region of GPs are realized. For an off-ramp bottleneck, at a given flow rate q_{in} , the larger the percentage of vehicles that leave the main road to the off-ramp η is, the larger the bottleneck strength.

At $\eta = 100\%$, all vehicles leave the main road to the off-ramp. Thus, if the lane number on the main road is larger than the lane number of the off-ramp, then at $\eta = 100\%$, the off-ramp bottleneck can be considered a merge bottleneck caused by the lane reduction.

It can be seen from the diagrams of congested patterns (Fig. 56) that when the flow rates (and/or η for an off-ramp bottleneck) change, complex transformations between various SPs as well as between SPs and GPs occur. The latter is more probable under weak congestion conditions within the GPs. Moreover, there are also regions in these diagrams in which SPs and GPs are metastable patterns. This means that pattern transformation can also occur at given flow rates (and a given η) as a result of one of the local phase transitions (e.g., $S \rightarrow F$, $S \rightarrow J$, or $J \rightarrow S$ transitions) within an initial congested pattern (for more detail, see Kerner 2004b).

It should be noted that theoretical results discussed in sections “Congested Patterns at Road Bottlenecks” and “Diagram of Congested Patterns at Isolated Bottlenecks” are taken from simulations of microscopic three-phase traffic flow models of Kerner and Klenov (2002, 2003), Kerner et al. (2002). Recently other various traffic flow models that incorporate some of the hypotheses of the three-phase theory have been developed (Davis 2004; Jiang and Wu 2004; Lee et al. 2004; Laval 2007; Jiang et al. 2007; Gao et al. 2007; Kerner and Klenov 2006; Zeng 2017). Features of congested patterns that these models exhibit (Jiang and Wu 2004; Lee et al. 2004; Laval 2007; Jiang et al. 2007; Gao et al. 2007; Davis 2006a, b, 2007; Jiang and Wu 2005, 2007; Wang et al. 2007; Pottmeier et al. 2007; Kerner and Klenov 2006; Zeng 2017) are similar to those found earlier in Kerner and Klenov (2002, 2003) and Kerner et al. (2002).

Complex Congested Patterns and Pattern Interaction at Adjacent Bottlenecks

When there are two or more adjacent bottlenecks, there can be many new spatiotemporal phenomena in congested traffic. In particular, we consider the following empirical spatiotemporal features of congested patterns (Kerner 2002b, 2004a, b):

- (i) Induced $F \rightarrow S$ transitions caused by upstream congested pattern propagation. A congested pattern, which occurs at the downstream bottleneck, can induce another

congested pattern at the upstream bottleneck where free flow is realized before.

- (ii) The catch effect. The catch of synchronized flow at an effectual bottleneck with the subsequent formation of a new congested pattern at this bottleneck.
- (iii) The occurrence of expanded congested patterns (EPs) where synchronized flow, which has initially appeared at the downstream bottleneck, affects the upstream bottleneck.
- (iv) The influence of foreign wide moving jam propagation on a congested pattern at the upstream bottleneck. These foreign wide moving jams occur downstream of the congested pattern within a GP at a downstream bottleneck.
- (v) An intensification of downstream congestion due to the onset of upstream congestion.

Induced F→S Transition with Catch Effect

In the empirical example shown in Fig. 44, when the MSP has not yet reached on-ramp bottleneck 1, free flow is there. After the MSP is caught at the bottleneck, synchronized flow occurs at on-ramp bottleneck 1. This synchronized flow continues for a long time (more than 2 hours) at on-ramp bottleneck 1. The upstream front of synchronized flow is now localized at some finite distance upstream of on-ramp bottleneck 1. This means that after the MSP has induced another SP at the bottleneck, instead of MSP propagation through the bottleneck, the MSP is caught at the bottleneck. Thus, the catch of the MSP at the bottleneck causes an induced speed breakdown at the bottleneck (these two effects are labeled “catch effect and induced F→S transition” in Fig. 44).

In simulations with microscopic three-phase traffic flow models, both induced traffic breakdown and catch effect can be found as observed in empirical data (Kerner 2004b). When two adjacent bottlenecks exist on a highway and either a WSP or MSP appears at the downstream bottleneck, then the associated SP can be caught at the upstream bottleneck with the subsequent formation of a new congested pattern upstream of the upstream bottleneck (Fig. 58).

For example, when the MSP shown in Fig. 58a reaches the upstream bottleneck, this

synchronized flow causes an induced traffic breakdown at this upstream bottleneck. Synchronized flow is localized at some distance from the bottleneck, rather than the upstream front of this synchronized flow propagating further upstream continuously over time. This means that the initial MSP is caught at the upstream bottleneck. This catch effect causes the induced traffic breakdown at the upstream bottleneck.

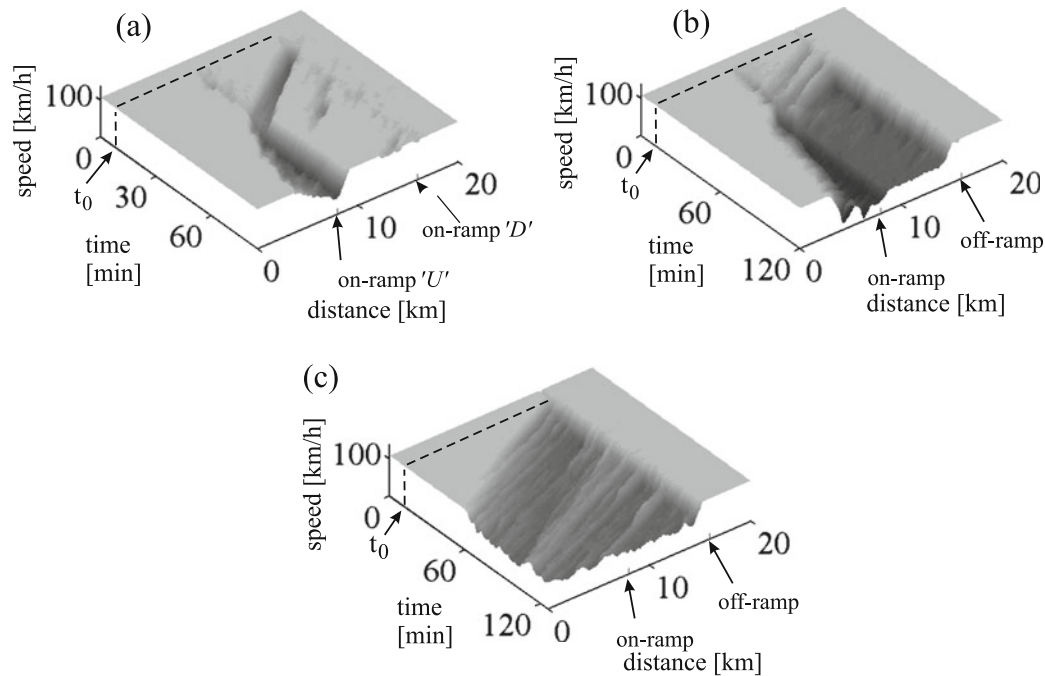
In another example, the upstream front of the WSP shown in Fig. 58b causes both induced traffic breakdown and catch effect after the upstream front of the WSP reaches the upstream bottleneck. Indeed, we can see that the initial WSP is caught at the upstream bottleneck. This catch effect causes also the induced traffic breakdown at the upstream bottleneck.

Expanded Congested Traffic Patterns (EPs)

An expanded congested pattern (EP) is defined as a congested traffic pattern whose synchronized flow affects at least two adjacent bottlenecks (Kerner 2002b, 2004b). A typical empirical example of an EP is shown in Fig. 42a, c: the WSP that has occurred initially at the downstream off-ramp bottleneck exists only for a relatively short time, which is determined by the propagation of the upstream front of the WSP to the upstream on-ramp bottleneck. In this case, the lifetime of the WSP is about 20 min. After synchronized flow of the WSP is caught at the upstream bottleneck, a new congested pattern is formed. Synchronized flow in this pattern covers both the off-ramp and on-ramp bottlenecks. Thus, in accordance with EP definition, this congested pattern is an EP.

If the above empirical congested pattern (Fig. 42) is considered during a considerably longer time interval and on the whole freeway section on which measurements are available, we find a typical complex EP with several wide moving jams propagating through the EP and two upstream on-ramp bottlenecks (on-ramp bottlenecks 1 and 2 in Fig. 59) while maintaining the mean velocity for the downstream jam front that are the same for these different jams (jams labeled by numbers 1, 2, ..., 6 in Fig. 59).

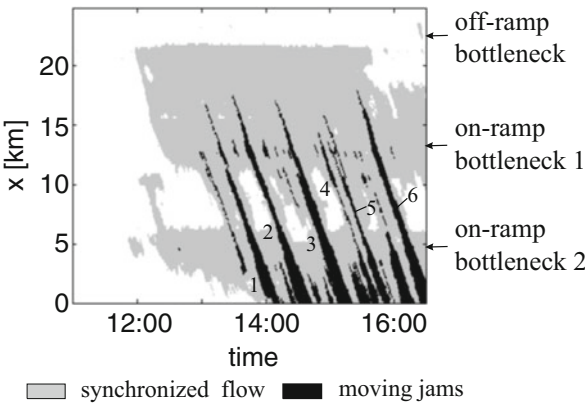
What is typical in this EP? Firstly, synchronized flow affects several effectual bottlenecks. Secondly,



Spatiotemporal Features of Traffic Congestion, Fig. 58 Numerical simulations of induced pattern formation with the catch effect (a, b) and expanded congested patterns (EPs) (b, c). (a, b) Catch effect of MSP (a) and

WSP (b) with induced formation of a new congested pattern at the upstream bottleneck. (Adapted from Kerner 2004b)

Spatiotemporal Features of Traffic Congestion, Fig. 59 Empirical example of an expanded congested pattern (EP). Free flow (white), synchronized flow (gray), moving jams (black). (Adapted from Kerner 2004b)



in addition to wide moving jams that have emerged between the off-ramp bottleneck and on-ramp bottleneck 1 (Fig. 59), new pinch regions in synchronized flow upstream of the on-ramp bottlenecks (on-ramp bottlenecks 1 and 2 in Fig. 59) are formed within which new narrow moving jams emerge spontaneously. Many of these narrow moving

jams, however, are often suppressed by the initially wide moving jams propagating through the pinch region. An explanation of the wide moving jam suppression effect can be found in Sect. 7.6.3 of the book Kerner (2004b). Thirdly, upstream and downstream of wide moving jams, either synchronized flow or free flow can be formed within the

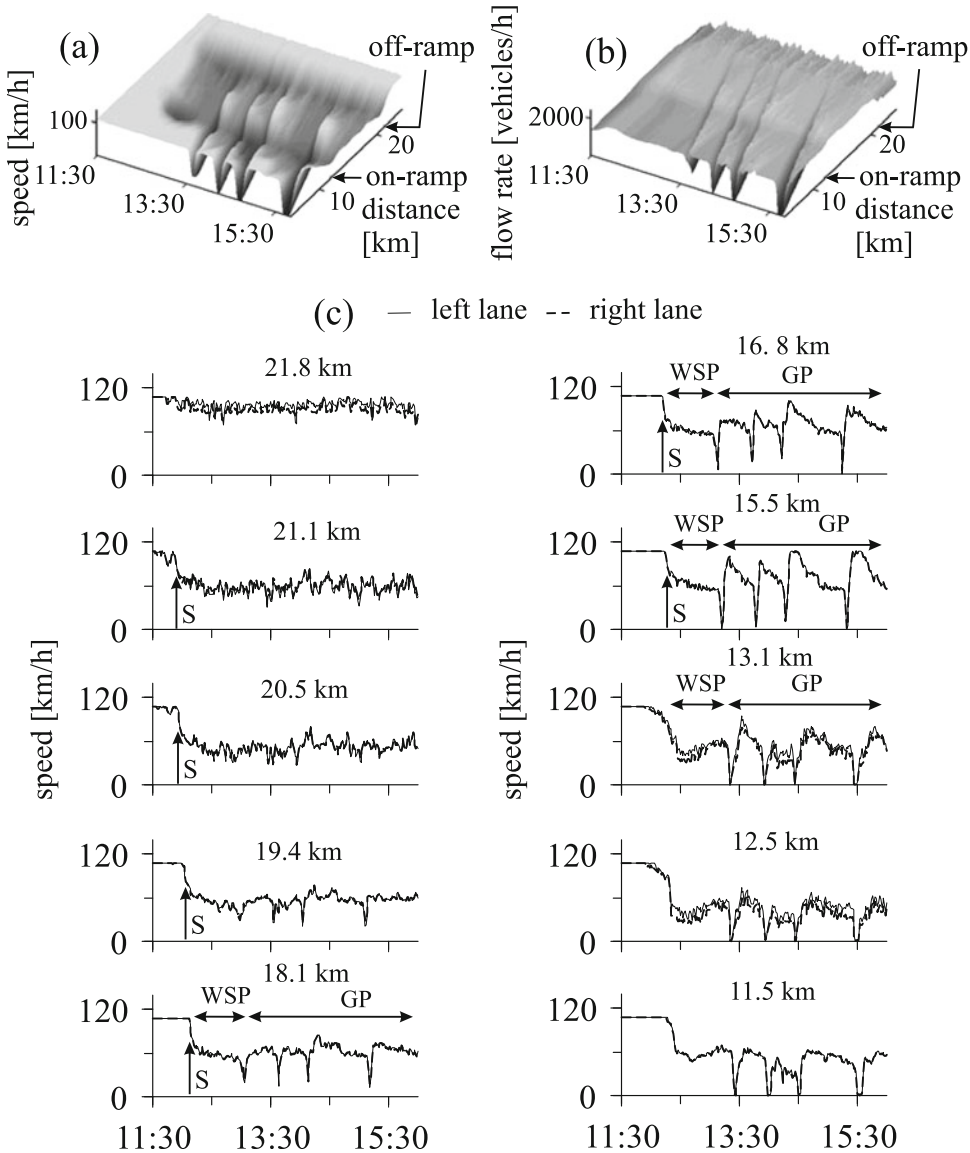
same EP. Note that much more complex EPs can be observed, when effectual bottlenecks are located at short distances to each other (see Chap. 2 in Kerner 2004b).

Simulations of a part of this EP (Fig. 60) made with the use of empirical road characteristics and flow rates are both qualitative and quantitative similar to the empirical EP (Figs. 48 and 59).

Foreign Wide Moving Jams

A foreign wide moving jam is a wide moving jam, which has initially occurred downstream of an effectual bottleneck and propagates through synchronized flow of another congested pattern upstream of this bottleneck.

We have mentioned above that wide moving jams in Fig. 59, which have initially emerged



Spatiotemporal Features of Traffic Congestion, Fig. 60 Simulations of a part of empirical EP shown in Figs. 48 and 59: (a, b) Speed (a) and flow rate (b) in space

and time. (c) Time-functions of speed at different locations. (Adapted from Kerner et al. 2006a)

upstream of the downstream off-ramp bottleneck, propagate through the upstream on-ramp bottleneck 1 while maintaining the velocity of the downstream jam front. These wide moving jams subsequently propagate through the next upstream on-ramp bottleneck 2 as well.

Upstream of the on-ramp bottlenecks 1 and 2 in Fig. 59, there are synchronized flows of other congested patterns formed. Thus, the wide moving jams labeled by numbers 1, 2, ..., 6 in Fig. 59 that initially emerged downstream of on-ramp bottleneck 1 are *foreign* wide moving jams when they propagate through synchronized flows upstream of on-ramp bottleneck 1 or the upstream of on-ramp bottleneck 2 (Kerner 2004b). An example of numerical simulations of foreign wide moving jam propagation is shown in Fig. 60.

A foreign wide moving jam can exert a considerable influence on other moving jams that are just emerging in synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam.

It can turn out that in addition to foreign wide moving jams, new wide moving jams can emerge in the pinch region of synchronized flow upstream of the upstream bottleneck. As a result, a “united” sequence of wide moving jams is formed from the joining of the foreign wide moving jams and the wide moving jams that have emerged in synchronized flow upstream of the upstream bottleneck.

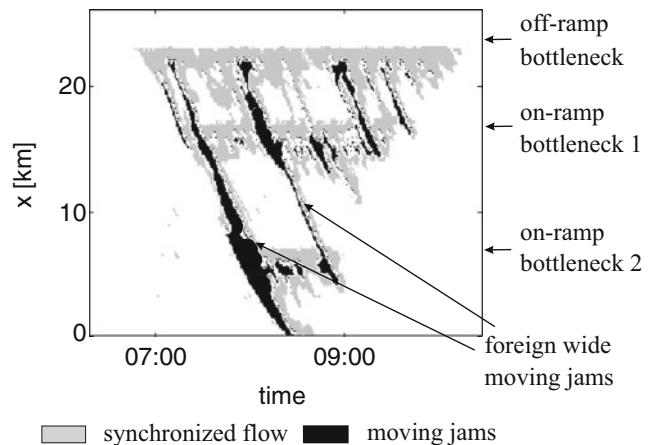
It should be noted that when the distances between adjacent effectual bottlenecks are large enough, spatially separated regions of synchronized flow can occur between adjacent effectual bottlenecks where the onset of congestion is realized. Often pinch regions are formed in these synchronized flows, leading to GP emergence. These GPs at different adjacent bottlenecks have spatially separated synchronized flow regions. We call these GPs “spatially separated” GPs (Kerner 2004b).

An empirical example is shown in Fig. 61, where two spatially separated GPs occur at adjacent effectual on-ramp bottlenecks (on-ramp bottleneck 1 and on-ramp bottleneck 2). Synchronized flow upstream of the GP at on-ramp bottleneck 1 does not cover on-ramp bottleneck 2, i.e., no EP occurs between them. In other words, a congested pattern can consist of several spatially separated GPs, if there is enough distance between adjacent effectual bottlenecks (distances should usually be more than 4–5 km).

There can also be an intermediate case. An example can be seen between downstream bottleneck (off-ramp bottleneck) and upstream bottleneck (on-ramp bottleneck 1) in Fig. 61. We find that during some time intervals, synchronized flow from the downstream GP at the off-ramp bottleneck covers the on-ramp bottleneck 1, i.e., EPs appear. During other time intervals, two spatially separated GPs at adjacent bottlenecks (off-ramp bottleneck and on-ramp bottleneck 1) are realized.

Spatiotemporal Features of Traffic Congestion,

Fig. 61 Empirical example of foreign wide moving jam propagation. Free flow (white), synchronized flow (gray), and moving jams (black) in space and time. (Adapted from Kerner 2004b)



In Fig. 61, several wide moving jams have emerged upstream of the downstream off-ramp bottleneck. Some of the wide moving jams propagate through adjacent on-ramp bottlenecks (on-ramp bottleneck 1 and on-ramp bottleneck 2) while maintaining the velocity of the downstream jam fronts. These wide moving jams are foreign wide moving jams when they are upstream of adjacent on-ramp bottlenecks (on-ramp bottleneck 1 and on-ramp bottleneck 2), where other moving jams are emerging. Thus, these GPs have only spatially separated regions of synchronized flow. In contrast to the synchronized flow regions, wide moving jams of the GP at the downstream bottleneck propagate through the GP at the upstream bottlenecks. As already mentioned, this foreign jam propagation can considerably influence on moving jam emergence at the upstream bottleneck. This occurs usually when spatially separated GPs appear. Thus, due to foreign wide moving jam propagation, spatially separated GPs can form a complex congested pattern on the freeway (Fig. 61).

Intensification of Downstream Congestion Due to Upstream Congestion

It can be expected that upstream congestion (congested pattern formation at an upstream bottleneck) should tend to a reduction in downstream congestion. This is because at the same traffic demand, the flow rate in free flow downstream of the congested bottleneck (discharge flow rate) is usually lower than the initial flow rate in free flow at the same freeway location before a

congested pattern at the upstream bottleneck has been formed.

Nevertheless in some cases, rather than a *reduction* in downstream congestion, an *intensification* of downstream congestion can occur due to upstream congestion (Kerner and Klenov 2003; Kerner 2004b). This is the result of complex non-linear interactions among congested patterns occurring at *different spatially separated* adjacent effectual bottlenecks.

To discuss an example of this effect, we consider a road with two on-ramp bottlenecks (Fig. 62). There are downstream and upstream on-ramp bottlenecks labeled by '*D*' and '*U*', respectively. The initial flow rate in free flow upstream of on-ramp '*D*'

$$q_{in}^{(down)} = q_{in} + q_{on}^{(up)} \quad (8)$$

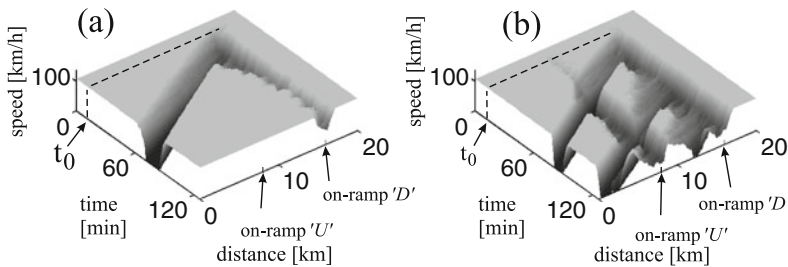
is the same in Fig. 62a, b. However, in Fig. 62a, the flow rate

$$q_{on}^{(up)} = 0, \quad (9)$$

whereas in Fig. 62b, the flow rate

$$q_{on}^{(up)} > 0. \quad (10)$$

The equality of the flow rate $q_{in}^{(down)}$ (8) in Fig. 62a, b is realized due to a larger flow rate q_{in} in free flow upstream of the upstream bottleneck in Fig. 62a in comparison with the flow rate q_{in} in Fig. 62b.



Spatiotemporal Features of Traffic Congestion, Fig. 62 Simulations of intensification of upstream congestion due to downstream congestion. Speed on the main

road in space and time: (a, b) DGP (a) transforms into a GP (b). (Adapted from Kerner and Klenov 2003; Kerner 2004b)

An initial DGP at the on-ramp ' D' ', where *only one* wide moving jam emerges (Fig. 62a), transforms into a GP at this bottleneck where an *uninterrupted sequence* of wide moving jams emerges (Fig. 62b).

To explain this intensification of downstream congestion, note that after the wide moving jam in the DGP (Fig. 62a) emerges, the current flow rate upstream of the on-ramp ' D' ', $q_{in}^{(down)}$, has dropped to q_{out} :

$$q_{in}^{(down)} = q_{out}. \quad (11)$$

This decrease in the flow rate $q_{in}^{(down)}$ occurs because vehicles merging onto the main road from the on-ramp ' U' ' must slow down when they reach the wide moving jam. At the flow rate $q_{in}^{(down)}$ (11), the GP cannot exist upstream of the on-ramp ' D' ' at the chosen flow rate $q_{on}^{(down)}$ to the on-ramp ' D' ' (Fig. 62a).

In Fig. 62b, an LSP is induced at the on-ramp ' U' ' after the wide moving jam from a DGP (downstream congestion) has reached the on-ramp ' U' '. After the wide moving jam has passed the on-ramp ' U' ', there are no wide moving jams between the on-ramps ' D' ' and ' U' '. For this reason, vehicles merging onto the main road from the on-ramp ' U' ' cause an increase in the flow rate $q_{in}^{(down)}$ to the value that is much larger than $q_{in}^{(down)} = q_{out}$ (11). This leads to subsequent wide moving jam emergence at the on-ramp ' D' ' and so on. As a result, a GP is formed at the on-ramp ' D' ' (Fig. 62b).

The mean period of moving jam emergence in the GP is determined by the time of wide moving jam propagation to the on-ramp ' U' ' (we neglect here the travel time in free flow). Before a wide moving jam has passed, the on-ramp ' U' ' condition (11) is satisfied, and no moving jam can emerge at the given flow rate $q_{on}^{(down)}$ at the on-ramp ' D' '.

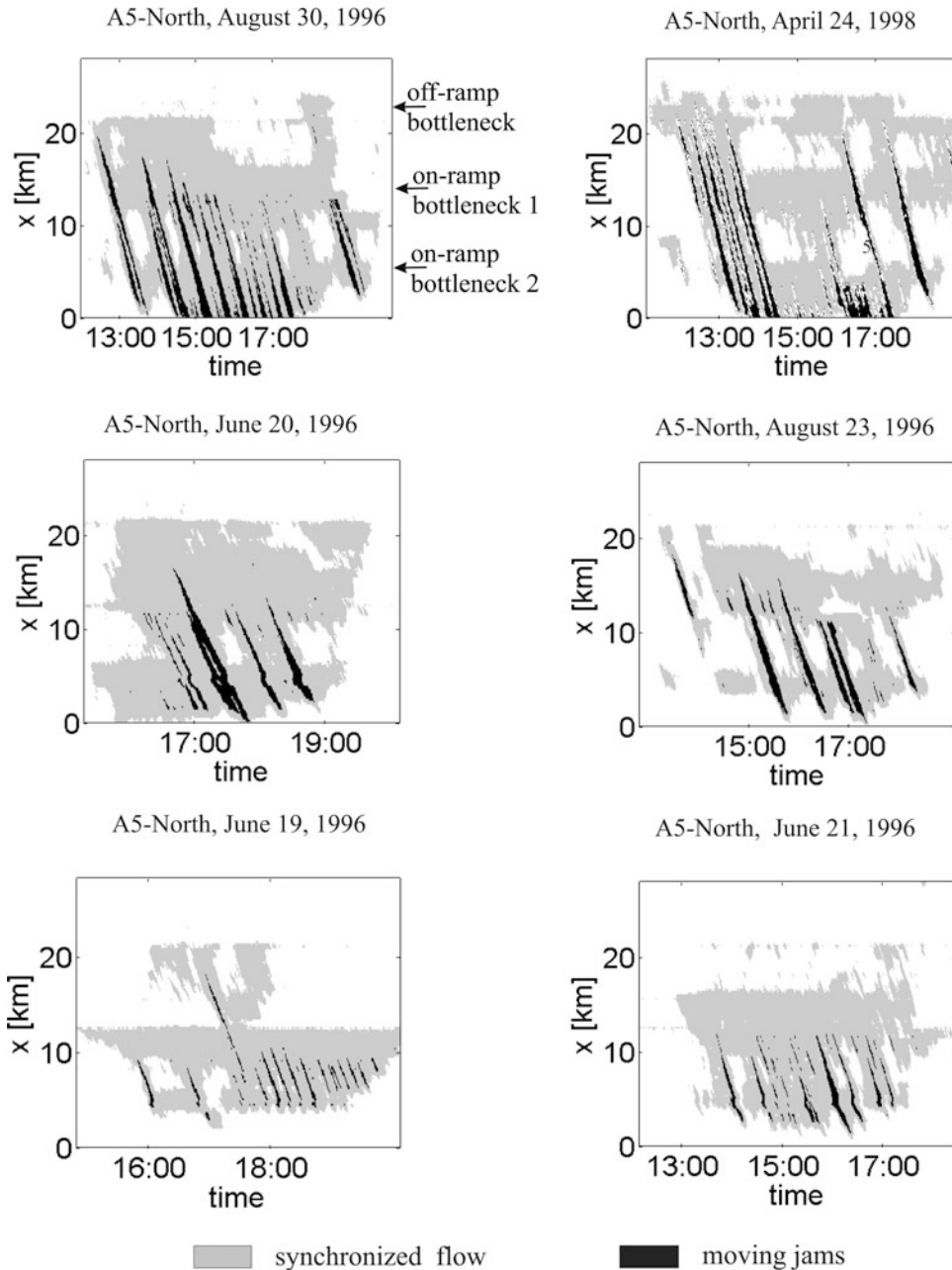
Reproducible and Predictable Congested Patterns

It has been found that for each effectual bottleneck or each set of several adjacent effectual bottlenecks, where congested patterns occur, the spatiotemporal structure of congested patterns exhibits predictable, i.e., characteristic, unique, and

reproducible, features (Kerner 2004b). These predictable and reproducible pattern features can be almost *the same* for different days and years. They can also persist over a large range of the flow rates (traffic demand) at which the patterns exist. The predictable and reproducible pattern features remain in traffic flows with very different driver behavioral characteristics and different vehicle parameters. The predictable and reproducible pattern features can be used in the forecast of congested patterns at freeway bottlenecks. Overviews of some of the reproducible congested pattern features can be seen in Fig. 63.

There are pattern features that have a probabilistic nature. This means that in different realizations (days) congested pattern features can be different for the same traffic control parameters (the percentage of long vehicles, weather, other road conditions) and traffic demand. However, we can identify different "degrees of predictability" for different pattern features. Some "deterministic" pattern features and pattern features with a high "degree of predictability" are as follows:

1. Traffic breakdown (F→S transition) occurs mostly at the same effectual freeway bottlenecks on different days and years.
2. In contrast to synchronized flow, which occurs mostly at bottlenecks, wide moving jam emergence is usually observed at the upstream boundary of the pinch region of synchronized flow that is at a finite distance upstream of bottlenecks.
3. Wide moving jams propagate through any complex state of traffic and through bottlenecks at the mean velocity of the downstream jam front.
4. If free flow emerges in the wide moving jam outflow, then the flow rate in this jam outflow is considerably lower than the maximum possible flow rate in free flow.
5. The catch effect: a local synchronized flow region reaching a bottleneck is caught at the bottleneck, when free flow at the bottleneck is in a metastable state with respect to traffic breakdown.
6. If a narrow moving jam is very close to the downstream front of a foreign wide moving



Spatiotemporal Features of Traffic Congestion, Fig. 63 Empirical examples of predictable and reproducible patterns. Overviews of free flow (white), synchronized

flow (gray), and moving jams (black) in space and time. The effectual bottlenecks are the same as those in Fig. 59. (Adapted from Kerner 2004b)

jam, the foreign jam suppresses the growth of this narrow jam.

Some traffic pattern features with a middle “degree of predictability” are as follows:

- (i) The instant of spontaneous traffic breakdown at an effectual bottleneck at the same time depends on upstream flow rates. This reveals the probabilistic nature of an $F \rightarrow S$ transition.

- (ii) For a bottleneck or for a set of adjacent effectual bottlenecks that are close to one another there is a certain type of spatiotemporal congested pattern that appears with the highest probability, given the same traffic demand.

Some pattern features with a low “degree of predictability” are as follows:

- (a) Whether a spontaneous or an induced traffic breakdown occurs, leading to synchronized flow at an effectual bottleneck
- (b) The instant of wide moving jam emergence in synchronized flow

Congested Patterns at Heavy Bottlenecks

In general, the larger the bottleneck influence on traffic flow (called the bottleneck strength) is, the smaller the average flow rate $q^{(\text{cong})}$ within a congested traffic pattern upstream of a bottleneck. As in section “[Diagram of Congested Patterns at Isolated Bottlenecks](#),” the time interval of the averaging of the flow rate $q^{(\text{cong})}$ is assumed to be considerably longer than time distances between any moving jams within a congested pattern. Empirical data show that the flow rate $q^{(\text{cong})}$ within GPs occurring at usual bottlenecks like on- and off-ramp bottlenecks discussed above in this article, which is equal to the average flow rate in the pinch region $q^{(\text{pinch})}$, is approximately within a range

$$q^{(\text{cong})} = 1100 - 1700 \text{ vehicles}/(\text{h lane}). \quad (12)$$

In contrast with the usual bottlenecks, due to, for example, bad weather conditions or accidents, heavy bottlenecks can occur. A heavy bottleneck exhibits a much larger influence on traffic (larger bottleneck strength). This leads to the limitation of the flow rate $q^{(\text{cong})}$ to very small values, sometimes as low as zero. Results of a theory of traffic congestion at heavy bottlenecks developed in Kerner (2007, 2008) are discussed in this section.

In the case of a heavy highway bottleneck, we can expect new interesting physical phenomena

associated with complexity of traffic congestion. This expectation follows already from a qualitative consideration of the influence of the decrease in $q^{(\text{cong})}$ on a sequence of wide moving jams. Indeed, from empirical single-vehicle data studied in Kerner et al. (2006a, b), we can conclude that the flow rate $q^{(\text{blanks})}$ of low speed states associated with moving blanks within wide moving jams satisfies an approximate condition

$$q^{(\text{blanks})} \lesssim 600 \text{ vehicles}/(\text{h lane}). \quad (13)$$

The average flow rate within wide moving jams $q^{(\text{blanks})}$ is defined as a number of vehicles $N^{(\text{blanks})}$, which have passed a virtual road detector during time intervals of the propagation of n_J wide moving jams (it is assumed that $n_J \gg 1$) through the detector, divided by the sum of all jam durations:

$$q^{(\text{blanks})} = \frac{N^{(\text{blanks})}}{\sum_{i=1}^{n_J} \tau_J^{(i)}}. \quad (14)$$

In (14), $\tau_J^{(i)}$ is a duration of an i -th wide moving jam, i.e., the time interval between the downstream and upstream fronts of the jam i , while this jam propagates through the detector.

Between the wide moving jams, the average flow rate associated with non-interrupted flows in the jam outflows $q_{\text{out}}^{(J)}$ is larger than $q^{(\text{pinch})}$ (12), i.e., $q_{\text{out}}^{(J)}$ is considerably larger than $q^{(\text{blanks})}$ (13). Now we assume that due to bad weather conditions or an accident, a heavy highway bottleneck occurs with a large strength at which the flow rate in congested traffic upstream of the bottleneck satisfies condition

$$q^{(\text{cong})} = q^{(\text{blanks})}. \quad (15)$$

In this case, $q_{\text{out}}^{(J)}$ must reduce to $q^{(\text{blanks})}$ (13), i.e., the difference between flows within and outside wide moving jams disappears. As a result, at

$$q^{(\text{cong})} \leq q^{(\text{blanks})} \quad (16)$$

wide moving jams should merge into a mega-wide moving jam (mega-jam for short).

For an analysis of traffic congestion at heavy bottlenecks, the Kerner-Klenov stochastic three-phase traffic flow model of vehicular traffic on a two-lane road with identical drivers and vehicles (Kerner and Klenov 2002, 2003) considered in Kerner (2009a) has been used with the following model for a heavy bottleneck.

We assume that there is a section of the road of a length L_B within which due to an accident or bad weather conditions, drivers should increase a safe time headway τ_{safe} to the preceding vehicle as well as decrease the maximum speed to some v_B that is lower than the maximum speed in free flow: within the section, the safe time headway is equal to $\tau_{\text{safe}} = T_B > 1$ s that is longer than 1 s used in the model for vehicles moving outside this section. In accordance with the model, τ_{safe} determines a safe speed, which should not be exceeded by a driver; otherwise, the driver decelerates. As a result, within this section, drivers move with the mean time headway that is longer than the mean time headway outside the bottleneck section, and the longer it is, the longer $\tau_{\text{safe}} = T_B$. Therefore, the section with longer $\tau_{\text{safe}} = T_B$ acts as a bottleneck on the road. In the bottleneck model, each chosen value T_B defines a specific bottleneck: the larger the strength of this bottleneck is, the longer the T_B . In its turn, the longer the T_B , the smaller the $q^{(\text{cong})}$, i.e., the larger the flow rate limitation within traffic congestion caused by the bottleneck.

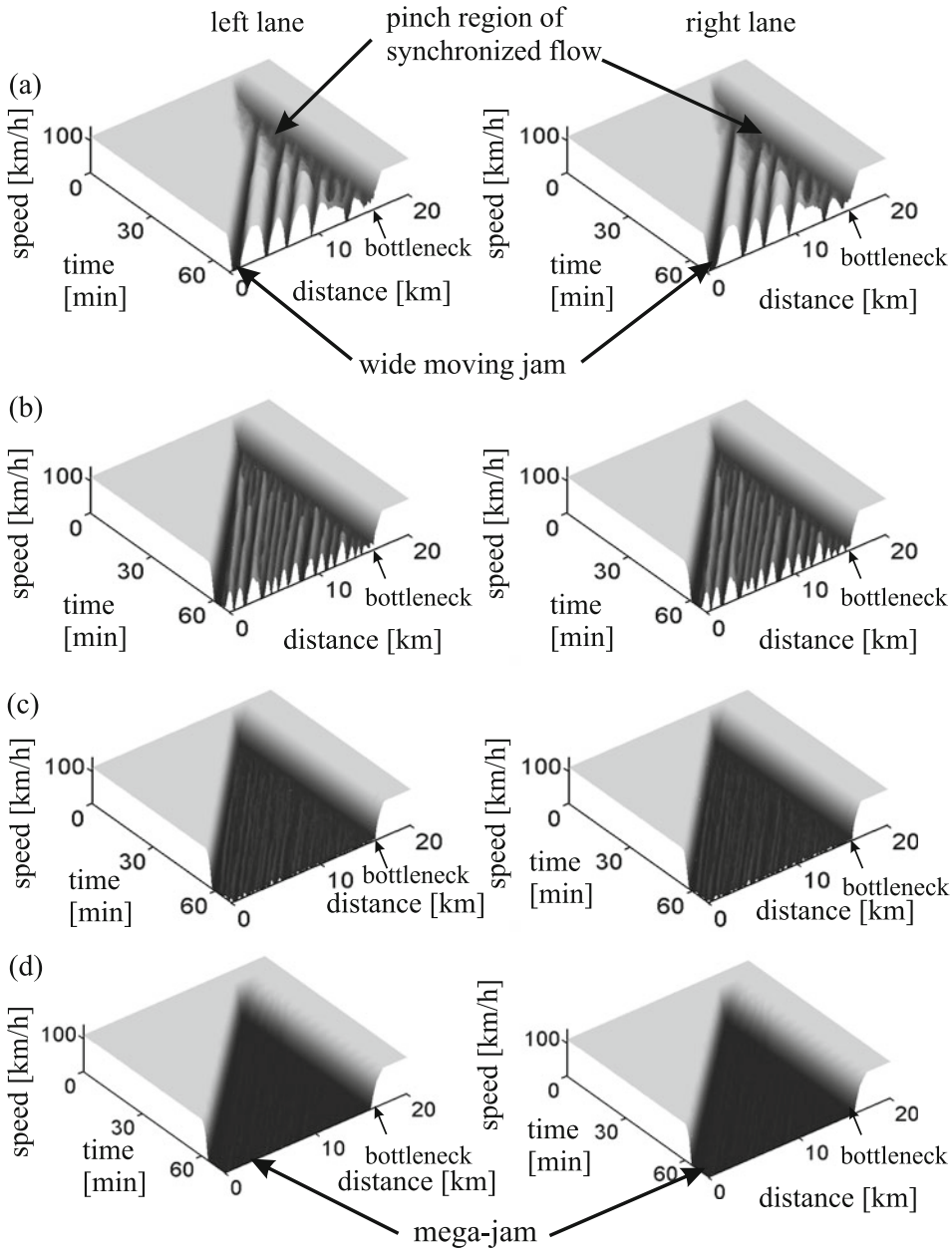
This model feature allows us to simulate a heavy bottleneck caused by bad weather conditions or accidents related to a long enough value T_B within the bottleneck. Under bad weather conditions, a road section with a long value T_B can be caused, e.g., by a poor view due to fog on the section or a much longer deceleration way needed by snow and ice on the section. If an accident occurs on a road, a road section with a long value T_B can be caused, e.g., by much narrower lane widths allowed for driving on the road section; the same effect can occur under heavy roadworks.

In simulations discussed below, we consider traffic phenomena occurring when the bottleneck strength increases gradually through the increase in T_B , when other bottleneck parameters are given constants (Fig. 64).

Evolution of Traffic Phases at Heavy Bottlenecks

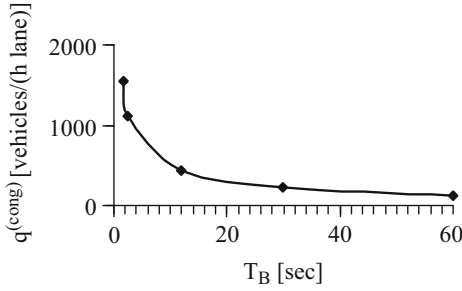
When the bottleneck strength is not large, i.e., T_B is chosen to be not very long ($T_B = 1.8$ s), then at a large enough flow rate q_{in} in free flow upstream of the bottleneck, firstly synchronized flow occurs spontaneously at the bottleneck. Then, the pinch region with a relatively large flow rate $q^{(\text{cong})} = q^{(\text{pinch})}$ is formed (Figs. 64a and 65). At the pinch region upstream boundary, wide moving jams emerge. Thus, we found the phenomena of regular GP formation (Figs. 64a and 66a, b) discussed in section “General Congested Patterns.” The pinch region width $L^{(\text{pinch})}(t)$ changes over time between about 1 and 2 km (Fig. 67a). $L^{(\text{pinch})}$ is defined as the distance between the upstream boundary of the bottleneck ($x = 16$ km) and the road location upstream at which a wide moving jam has just been identified through the use of the jam microscopic criterion (2). There is also a nearly constant frequency of $L^{(\text{pinch})}(t)$ oscillations associated with the maximum in the Fourier spectrum (Fig. 67b). Speed autocorrelation functions and associated Fourier spectra of speed time dependencies at shorter T_B show regular character of wide moving jam propagation (Fig. 66c, d). In other words, at such short values T_B , the bottleneck model discussed in this section “Congested Patterns at Heavy Bottlenecks” shows qualitatively the same phenomena of spatiotemporal congested patterns that occur at usual highway bottlenecks (sections “Congested Patterns at Road Bottlenecks” and “Microscopic Features of Traffic Phases in Congested Traffic”).

When T_B becomes longer and, therefore, the bottleneck strength increases, the flow rate $q^{(\text{cong})} = q^{(\text{pinch})}$ decreases (Fig. 65). However, if T_B remains a relatively small value ($1.8 < T_B < 3$ s), then as for other bottleneck types, we found the following GP features (some of them have already been discussed in section “General Congested Patterns”): the smaller the flow rate $q^{(\text{pinch})}$, the larger the frequency of narrow moving jam emergence within the pinch region, the lower the maximum (and the average) speed between wide moving jams upstream of the pinch region, and the smaller the mean pinch



Spatiotemporal Features of Traffic Congestion, Fig. 64 Simulated speed in space and time in the left (left) and right (right) road lanes at different T_B : $T_B = 1.8$ (a), 2.4 (b), 12 (c), and 60 s (d). $q_{in} = 1946$ vehicles/(h lane). The upstream boundary of bottleneck region of

the length $L_B = 300$ m is at $x = 16$ km; the maximum speed within this bottleneck region is $v_B = 60$ km/h. Resulting values of $q^{(cong)} = 1546$ (a), 1114 (b), 440 (c), and 127 (d) vehicles/(h lane). (Adapted from Kerner 2009a)



Spatiotemporal Features of Traffic Congestion, Fig. 65 Simulated average flow rate $q^{(congestion)}$ within congested patterns shown in Fig. 64 as a function of T_B . The averaging time interval for $q^{(congestion)}$ is 60 min. (Adapted from Kerner 2009a)

region width $L_{mean}^{(pinch)}$ (Figs. 64b and 67c, d, k, l). This can also be seen from a comparison of time dependencies of average speed within the region of wide moving jams for different T_B (Fig. 66a, e).

Qualitatively other phenomena are found when T_B further increases ($T_B > 3$ s) and the average flow rate $q^{(congestion)}$ decreases considerably (Figs. 64c, d and 65).

We find that there is a critical strength of a heavy bottleneck associated with $T_B = T_B^{(break)}$ that results in the critical flow rate $q^{(congestion)} = q_{break}^{(congestion)}$. When

$$q^{(congestion)} \leq q_{break}^{(congestion)} \quad (17)$$

related to $T_B \geq T_B^{(break)}$, then there are random time intervals when the pinch region disappears, i.e., $L^{(pinch)} = 0$ (Fig. 67e, g). This means that there are time instants at which there is no pinch region and wide moving jams emerge directly at the upstream boundary of the bottleneck, whereas for other time intervals, the pinch region appears again (Fig. 67e, g). We found that $q_{break}^{(congestion)} \approx 920$ vehicles/(h lane) ($T_B^{(break)} \approx 3$ s at the chosen bottleneck parameters L_B and v_B). Under condition (17), $L^{(pinch)}(t)$ becomes a non-regular time-function (Fig. 67e, g, i) whose Fourier spectrum is broader, the longer T_B , i.e., the smaller $q^{(congestion)}$ (Fig. 67f, h, j). Such a GP at an isolated heavy bottleneck can be considered the GP with a non-regular pinch region.

The more the bottleneck strength exceeds the critical bottleneck strength, the larger the difference $q_{break}^{(congestion)} - q^{(congestion)}$ and the longer the mean duration of time intervals within which the regular structure of the GP breaks and the pinch region disappears, and therefore, the smaller the mean length $L_{mean}^{(pinch)}$ of the pinch region (Fig. 67k, l). In contrast with regular time dependencies of the average speed within a sequence of wide moving jams (Fig. 66a–h), the time-functions of 1-min average speed at a larger bottleneck strength exhibit a non-regular behavior (Fig. 66i–l). This can be seen from speed autocorrelation functions and associated Fourier spectra of the average speed time dependencies (Fig. 66k, l).

When the bottleneck strength increases further, $L_{mean}^{(pinch)}$ decreases continuously (Fig. 67l). $L_{mean}^{(pinch)}$ reaches zero at a threshold bottleneck strength for the pinch region existence associated with $T_B = T_B^{(th)}$ that results in a threshold flow rate $q^{(congestion)} = q_{th}^{(congestion)}$. When

$$q^{(congestion)} \leq q_{th}^{(congestion)} \quad (18)$$

related to $T_B \geq T_B^{(th)}$, then the pinch region of GPs does not exist. At model parameters, the pinch region of a GP disappears fully at $q_{th}^{(congestion)} \approx 220$ vehicles/(h lane) ($T_B^{(th)} \approx 30$ s at the chosen values of L_B and v_B).

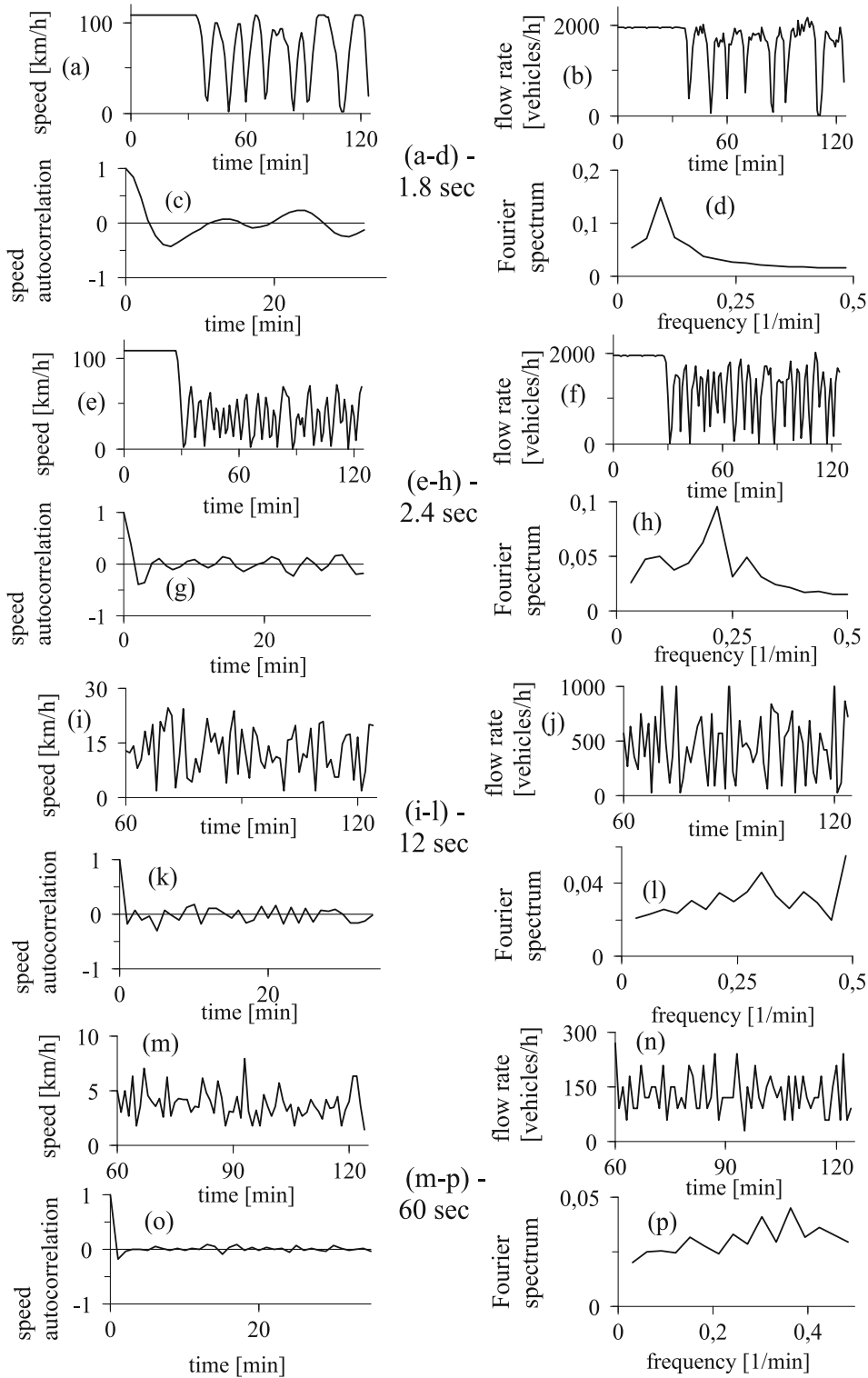
There is also another critical bottleneck strength, which we call the critical bottleneck strength for the mega-jam formation associated with $T_B = T_B^{(mega)}$ that results in the critical flow rate $q^{(congestion)} = q_{mega}^{(congestion)}$. When

$$q^{(congestion)} \leq q_{mega}^{(congestion)} \quad (19)$$

related to $T_B \geq T_B^{(mega)}$, then wide moving jams merge onto a mega-jam. Numerical simulations show that (19) is equivalent to (16), i.e.:

$$q_{mega}^{(congestion)} = q^{(blanks)}. \quad (20)$$

At model parameters, $q_{mega}^{(congestion)} \approx 130$ vehicles/(h lane) ($T_B^{(mega)} \approx 60$ s at the chosen values of L_B and v_B).

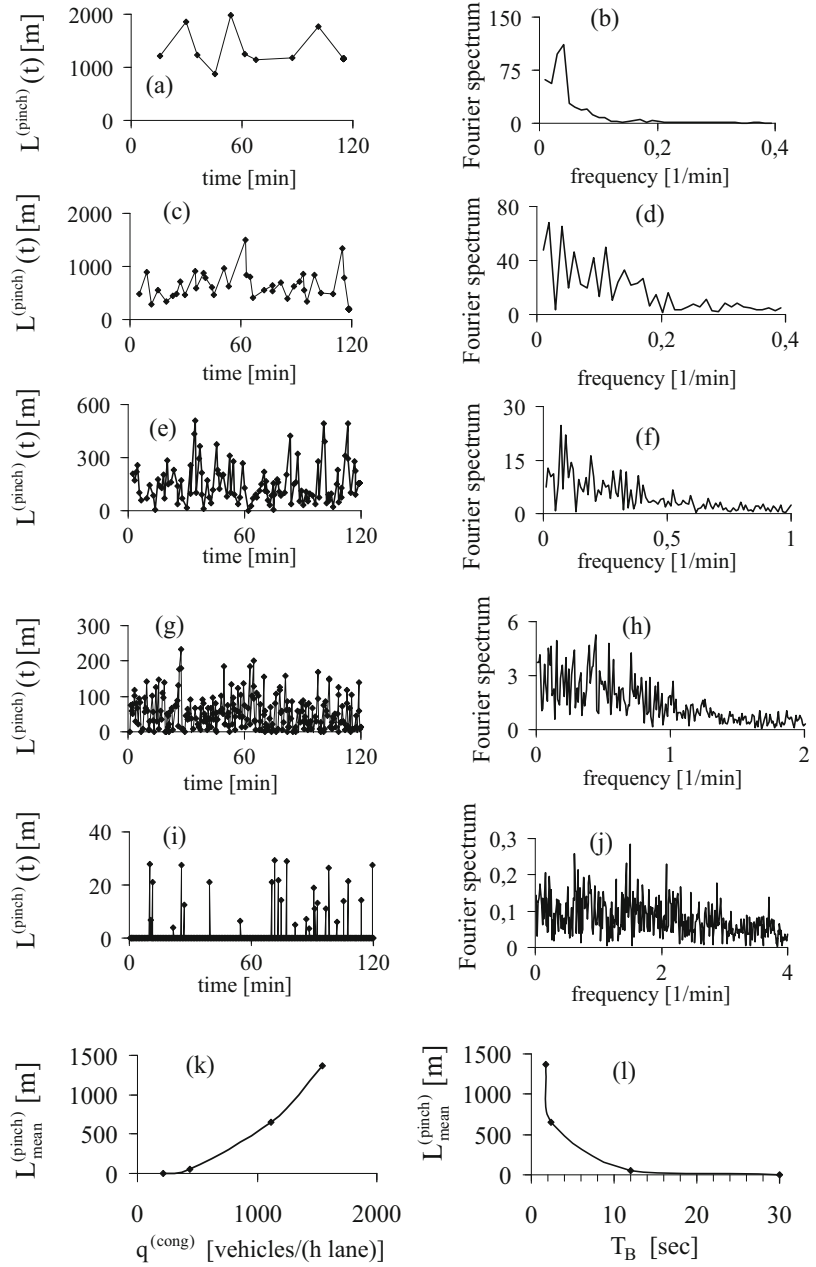


Spatiotemporal Features of Traffic Congestion, Fig. 66 Simulated characteristics of congested patterns shown in Fig. 64 related to location 10 km. Time-functions of speed (a, e, i, m), speed correlations (c, g, k, o), associated

Fourier spectra (d, h, l, p), and flow rate (b, f, j, n) for different $T_B = 1.8$ (a-d), 2.4 (e-h), 12 (i-l), 60 s (m-p). 1-min average data in the left lane. (Adapted from Kerner 2009a)

Spatiotemporal Features of Traffic Congestion,

Fig. 67 Simulations of pinch effect: $L^{(\text{pinch})}(t)$ (a, c, e, g, i), associated Fourier spectra (b, d, f, h, j) for congested patterns related to $T_B = 1.8$ (a, b), 2.4 (c, d), 6 (e, f), 12 (g, h), and 30 s (i, j). (k, l) $L_{\text{mean}}^{(\text{pinch})}(q^{(\text{cong})})$ (k) and $L_{\text{mean}}^{(\text{pinch})}(T_B)$ (l). (Adapted from Kerner 2009a)



We have also found that the critical bottleneck strength for the mega-jam formation is larger than the threshold bottleneck strength for the pinch region existence that results in the condition

$$q_{\text{mega}}^{(\text{cong})} < q_{\text{th}}^{(\text{cong})} \quad (21)$$

related to $T_B^{(\text{mega})} > T_B^{(\text{th})}$. Under condition (19), i.e., when all wide moving jams merge onto a mega-jam, traffic congestion upstream of an isolated heavy bottleneck cannot be considered as a GP any more. Thus, in accordance with (19), (20), if traffic breakdown has occurred at a bottleneck, then the condition

$$q^{(\text{cong})} > q^{(\text{blanks})} \quad (22)$$

is a *necessary* condition for GP existence at this bottleneck.

As for GPs with a very non-regular pinch region (Fig. 66i–l), the time-functions of 1-min average speed within the mega-jam exhibit a non-regular behavior (Fig. 66m); this can be seen from speed autocorrelation functions and associated Fourier spectra of the average speed time dependencies (Fig. 66o, p).

When $q^{(\text{cong})}$ becomes zero, because behind a road location the road is closed, the mega-jam transforms into a queue of *motionless* vehicles, which therefore is not associated with vehicular traffic. Nevertheless, there is a link between the queue and the mega-jam. If at a time instant we allow several vehicles to escape from this queue, then simulations show that motion of these vehicles results in wide moving jam occurrence: the downstream front of the jam separates moving vehicles escaping from the queue and vehicles standing within the queue upstream of the front. When the number of vehicles escaping from the initial queue decreases, the downstream jam front transforms into a moving blank(s) subsequently covered by vehicles standing in the queue. When a vehicle per a long enough time interval is allowed to escape from the mega-jam, as simulations show, a sequence of such moving blanks within this jam occur; these moving blanks exhibit, however, the dynamics specifically associated with the mega-jam (see section “[Microscopic Spatiotemporal Structure of Mega-jam](#)” below).

Thus, based on a study of a three-phase traffic flow model, we found that the complexity of traffic congestion at a heavy bottleneck caused, for example, by bad weather conditions or accidents, is associated with the phenomenon of random disappearance and appearance of the pinch region over time as well as with the phenomenon of the merger of wide moving jams into a mega-jam. When the bottleneck strength increases continuously, the mean length of the pinch region decreases also continuously up to zero; at such a heavy bottleneck, the pinch region does not exist any more. Beginning from a larger

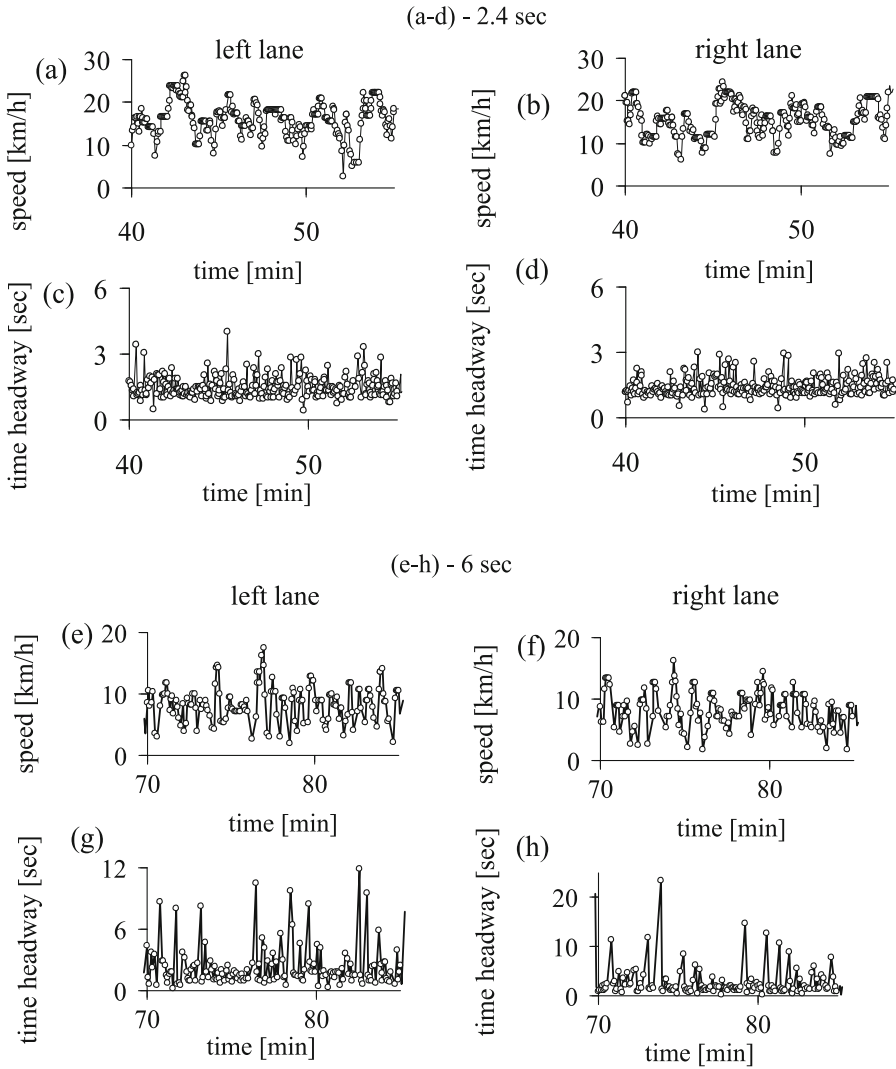
bottleneck strength, only the mega-jam survives in congested traffic upstream of the bottleneck. In this case, synchronized flow remains only within its downstream front separating free flow and congested traffic. These phenomena reveal the evolution of the traffic phases in congested traffic when heavy bottlenecks occur in highway traffic.

Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics

For GPs with a regular pinch region (condition (17) is not satisfied), we find results of section “[General Congested Patterns](#)” that at a small enough distance upstream of the bottleneck, there is the pinch region of non-interrupted synchronized flow (Fig. 68a, b) with short enough time headway for which criterion (2) is satisfied for *none* of time headway (Fig. 68c, d) (Kerner 2004b).

In contrast, under condition (17), when GPs with a non-regular pinch region occur, at the same small distance upstream of the bottleneck, some long time headway appears in the pinch region of the GP for which criterion (2) is satisfied (Fig. 68g, h). This means that the pinch region disappears during some time intervals and, therefore, during these time intervals, there are wide moving jams that emerge almost directly upstream of the bottleneck. This is related to the previous result that the pinch region width $L^{(\text{pinch})}$ of such a GP is nearly as small as zero at some random time instants (Fig. 67e, g). During other time intervals, there is synchronized flow with a short enough time headway for which criterion (2) is not satisfied (Fig. 68g, h). This is related to the conclusion of section “[Evolution of Traffic Phases at Heavy Bottlenecks](#)” that for the GPs with a non-regular pinch region at some random time instants, the pinch region disappears and appears again.

Single-vehicle data measured by a virtual road detector 6 km upstream of the bottleneck within the GPs with regular (Fig. 69a–d) and non-regular pinch regions (Fig. 69e–h) exhibits qualitatively the same and typical time dependencies of the speed and time headway for a sequence of wide



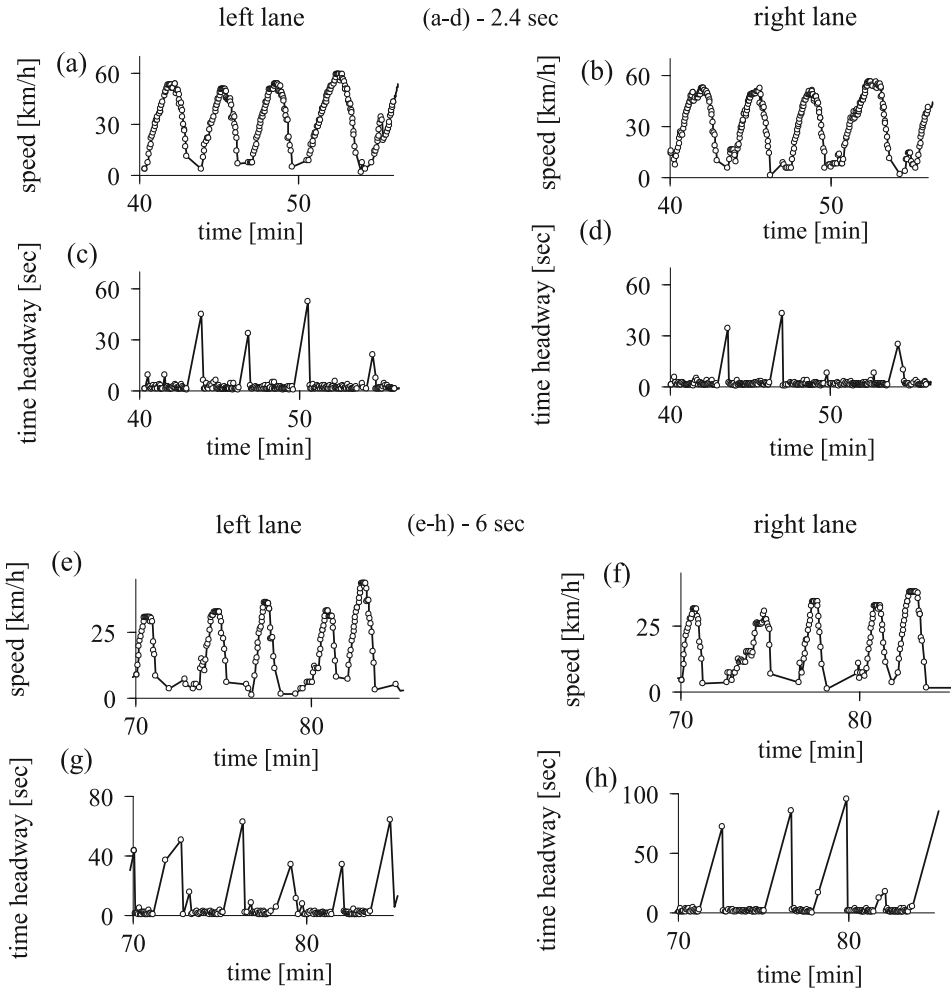
Spatiotemporal Features of Traffic Congestion, Fig. 68 Simulated single-vehicle data for speed (a, b, e, f) and time headway (c, d, g, h) measured by a virtual detector at location 15.8 km within congested traffic in the

left (figures left) and right lanes (right) related to different $T_B = 2.4$ (a–d) and 6 s (e–h). For (e–h) $q^{(\text{cong})} = 626$ vehicles/(h lane). (Adapted from Kerner 2009a)

moving jams of a GP (Kerner et al. 2006b). We see that there are moving jams within which the maximum time headway $\tau_{\max}$ between two vehicles within the jams is very long, i.e., traffic flow is interrupted, and criterion (2) is satisfied (Fig. 69c, d, g, h). Therefore, these jams are wide moving ones. There are long time intervals between these jams within which the speed is high (Fig. 69a, b, e, f) and vehicle time headway is small (Fig. 69c, d, g, h). These long regions between the

wide moving jams are related to non-interrupted traffic flow. Thus, at some distance upstream of the bottleneck, there is a sequence of wide moving jams that is characteristic for GPs.

In accordance with results of section “Microscopic Features of Traffic Phases in Congested Traffic,” a wide moving jam consists of alternations of flow interruption intervals and moving blanks within the jam. In a general case, for a wide moving jam, we can observe two or more



Spatiotemporal Features of Traffic Congestion, Fig. 69 Simulated single-vehicle data for speed (a, b, e, f) and time headway (c, d, g, h) measured by a virtual detector at location 10 km within congested traffic in the

left (figures left) and right lanes (right) related to different $T_B = 2.4$ (a–d) and 6 s (e–h). (Adapted from Kerner 2009a)

flow interruption intervals, i.e., two or more long time headway $\tau^{(k)}$, $k = 1, 2, \dots$ and each of them satisfies criterion (2):

$$I_s^{(k)} = \tau^{(k)} / \tau_{\text{del}}^{(\text{acc})}(0) \gg 1, \quad k = 1, 2, \dots, \quad (23)$$

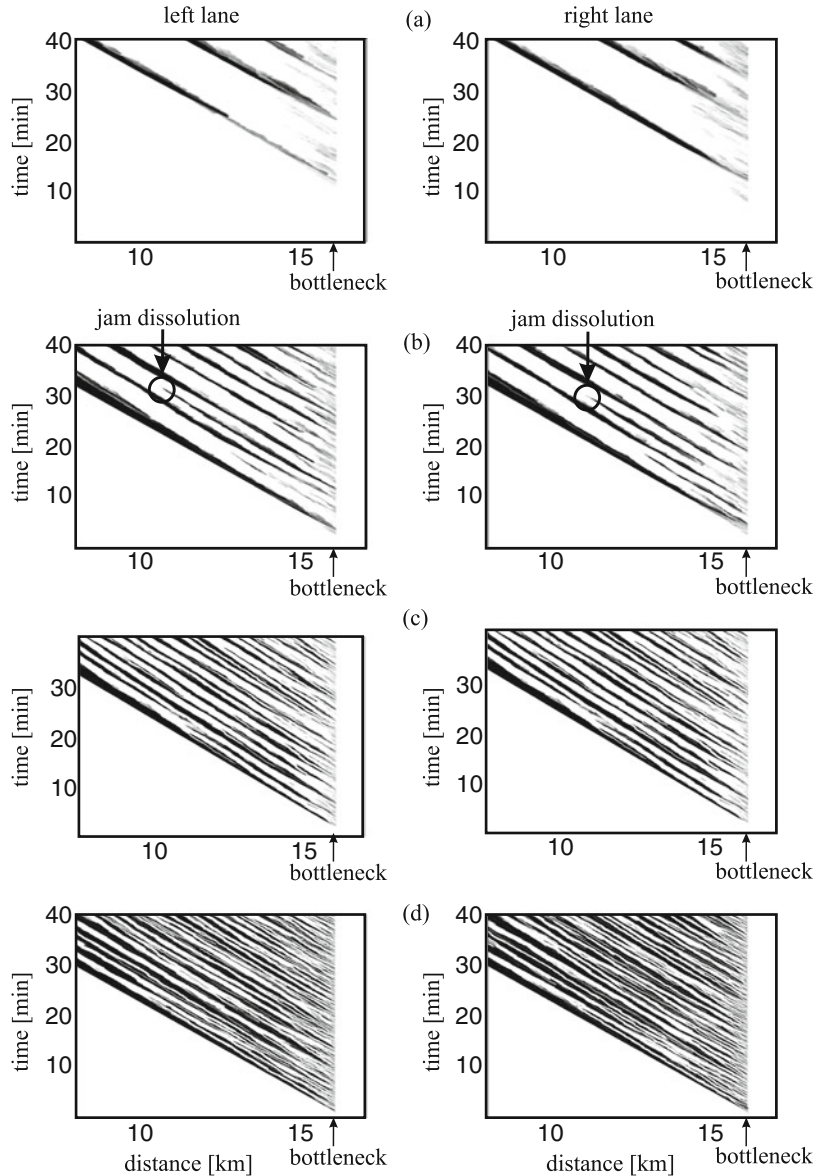
Between these flow interruption intervals, we find one or more vehicles that have passed the detector with speeds, which are considerably lower than the average speed both upstream and downstream of the wide moving jam. This low-speed vehicle motion is associated with moving blanks (section “Numerical Simulations of Moving

Blanks Within Wide Moving Jams”): the flow interruption intervals $\tau^{(k)}$, $k = 1, 2, \dots$ are separated by one or more moving blanks within the wide moving jam.

At low scales in time and space, both GPs with a regular (Fig. 70a, b) and non-regular pinch regions (Fig. 70c, d) exhibit *spatiotemporal microscopic structures* in which regions of lower speed alternate with regions of higher speed; in both cases, we find that these regions propagate upstream. At the first glance, we can see only that the mean frequency of low-speed regions increases with the increase in T_B , i.e., when the bottleneck

Spatiotemporal Features of Traffic Congestion,

Fig. 70 Simulated spatiotemporal microscopic structures of GPs with regular (a, b) and non-regular pinch regions (c, d): Single-vehicle speeds within GPs presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (left) and right (right) road lanes associated with speed distributions at different $T_B = 1.8$ (a), 2.4 (b), 4 (c), and 12 s (d). For (c) $q^{(\text{cong})} = 776$ vehicles/(h lane). The upstream boundary of the bottleneck region is at location $x = 16$ km (labeled “bottleneck”). (Adapted from Kerner 2009a)



strength increases; this result has already been mentioned in section “[Evolution of Traffic Phases at Heavy Bottlenecks](#).” However, if we consider the microscopic structures of GPs with a non-regular pinch region in larger scales in space and time (Figs. 71 and 72), we can see that there is a crucial qualitative difference between them and the microscopic structures of GPs with a regular pinch region (Fig. 70a, b).

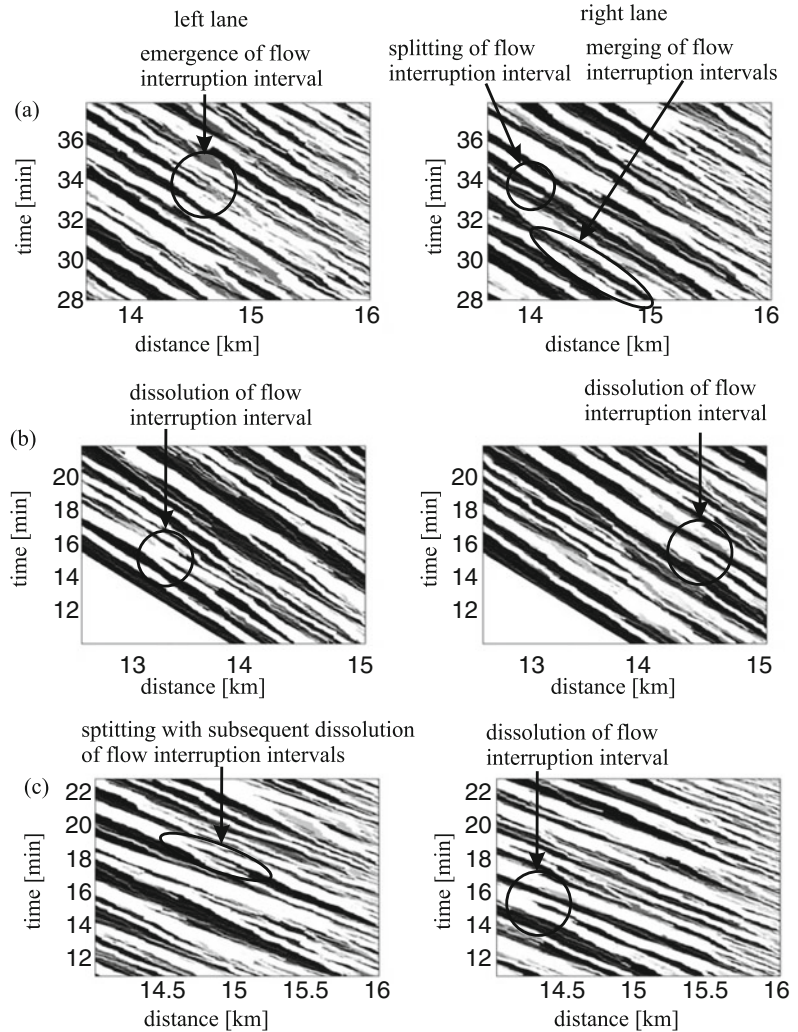
GPs with a regular pinch region exhibits a regular dynamics of wide moving jams discussed

in section “[General Congested Patterns](#),” when the jams propagate upstream of the pinch region on a homogeneous road (Fig. 70a, b): some of the wide moving jams can dissolve during their upstream propagation (a so-called jam suppression effect) (Kerner 2004b). An example of the effect of the dissolution of wide moving jams can be seen in Fig. 70b (labeled “jam dissolution”).

However, under condition (17), i.e., when a GP with a non-regular pinch region is realized, wide moving jams of the GP can exhibit complex and

Spatiotemporal Features of Traffic Congestion,

Fig. 71 Fragments of Fig. 70d in larger scales in time and space in the left (figures left) and right lanes (right). White regions (single-vehicle speeds are equal to or higher than 5.4 km/h) are related to synchronized flows and moving blanks. Black regions (single-vehicle speeds are equal to zero) are related to flow interruption intervals. $T_B = 12$ s. (Adapted from Kerner 2009a)



non-regular spatiotemporal dynamic behavior (Figs. 71 and 72). This non-regular jam dynamics is associated with the following effects (1–4):

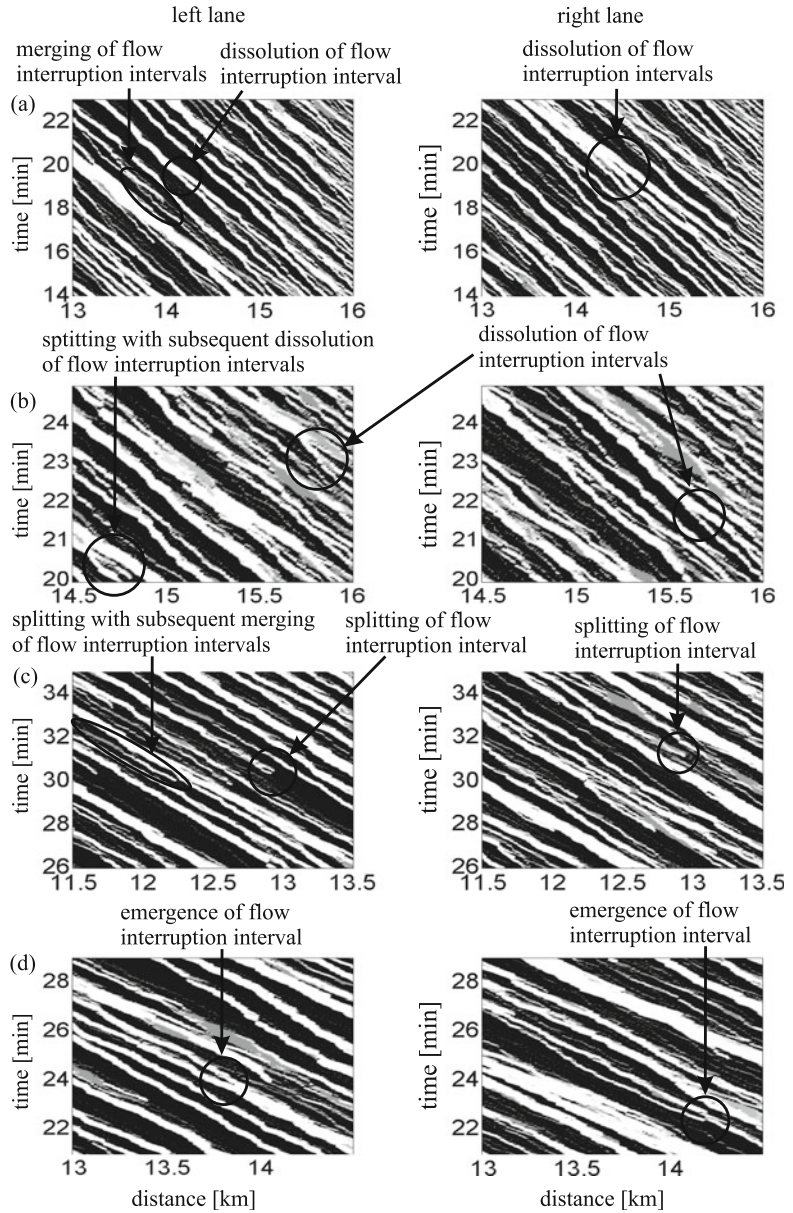
1. The effect of the *splitting* of a flow interruption interval onto two (or more) flow interruption intervals (Figs. 71a and 72c)
2. The effect of the *emergence* of a new flow interruption interval (Figs. 71a and 72d)
3. The effect of the *merging* of two (or more) traffic flow interruption intervals (Figs. 71a and 72a)
4. The effect of the *dissolution* of a flow interruption interval (Figs. 71b, c and 72)

These dynamic effects 1–4 occur spontaneously during upstream propagation of wide moving jams. In different lanes, these effects occur often independently of each other and at different road locations. A spatiotemporal competition of these effects and a diverse variety of the *initiating* and *resulting* dynamic effects determine a non-regular dynamic behavior of wide moving jams. The variety of the initiating and resulting effects associated with the effects 1–4 is as follows:

- (i) The splitting of a flow interruption interval within a wide moving jam can result from the emergence of a new region of moving blanks within the jam.

Spatiotemporal Features of Traffic Congestion,

Fig. 72 Simulated single-vehicle speeds within GP presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (left) and right lanes (right). White regions (single-vehicle speeds higher than 3.6 km/h) are related to synchronized flows and moving blanks. Black regions (single-vehicle speeds are equal to zero) are related to flow interruption intervals. $T_B = 30$ s. $q^{(\text{cong})} = 218$ vehicles/(h lane). (Adapted from Kerner 2009a)



- (ii) The splitting of a flow interruption interval within an initial wide moving jam can result in the splitting of the jam onto two (or more) wide moving jams; in the latter case, synchronized flow(s) (or free flow(s)) occurs that separates the emergent wide moving jams. This is related to an J→S (or J→F) transition(s) occurring within the initial jam.
- (iii) The effect of the emergence of a new flow interruption interval within a wide moving jam can result from the splitting of a region of moving blanks onto two (or more) regions of moving blanks separated by the emergent flow interruption interval.
- (iv) The effect of the emergence of a new flow interruption interval occurring within synchronized flow between two wide moving

jams can result in the emergence of a new wide moving jam. This is related to an S→J transition occurring in metastable synchronized flow between the jams.

- (v) The effect of the merging of two (or more) traffic flow interruption intervals can result from the dissolution of two (or more) regions of moving blanks within a wide moving jam without wide moving jam dissolution.
- (vi) The effect of the merging of two (or more) traffic flow interruption intervals can result in the merging of two (or more) wide moving jams.
- (vii) The effect of the dissolution of a traffic flow interruption interval within a wide moving jam can result from the merging of two (or more) regions of moving blanks within the jam.
- (viii) The effect of the dissolution of a traffic flow interruption interval can result in the dissolution of a wide moving jam.

The initiating and resulting effects (v)–(viii) decrease the mean frequency of regions of flow interruption intervals and regions of moving blanks within wide moving jams as well as the mean frequency of wide moving jams of a GP. In particular, the effects (vii) and (viii) occur very frequently close to the upstream boundary of the pinch region of GPs (see road locations between 15 and 16 km in Figs. 71a, c and 72a, b). This leads to a decrease in the mean frequency of wide moving jams, when the jams propagate upstream.

There can be diverse sequences of the effects 1–4 over time with a random and arbitrary sequence of the effects (i)–(viii). For example, the splitting of a flow interruption interval onto two intervals can be followed by the subsequent dissolution of these two intervals (Figs. 71c and 72b); the splitting of a flow interruption interval onto two interruption intervals can be followed by the subsequent merging of these two interruption intervals (Fig. 72c). The random and arbitrarily spatiotemporal sequence of the effects discussed can explain the non-regular spatiotemporal jam dynamics found in GPs with a non-regular pinch region. In turn, this non-regular spatiotemporal

jam behavior can also explain the result of section “[Evolution of Traffic Phases at Heavy Bottle-necks](#)” that we cannot distinguish a sequence of wide moving jams in 1-min average data (Fig. 66i–l): the microscopic non-regular jam dynamics is averaged in this 1-min average data leading to non-regular and nonhomogeneous speed distributions in which this real spatiotemporal jam dynamics *cannot be found*.

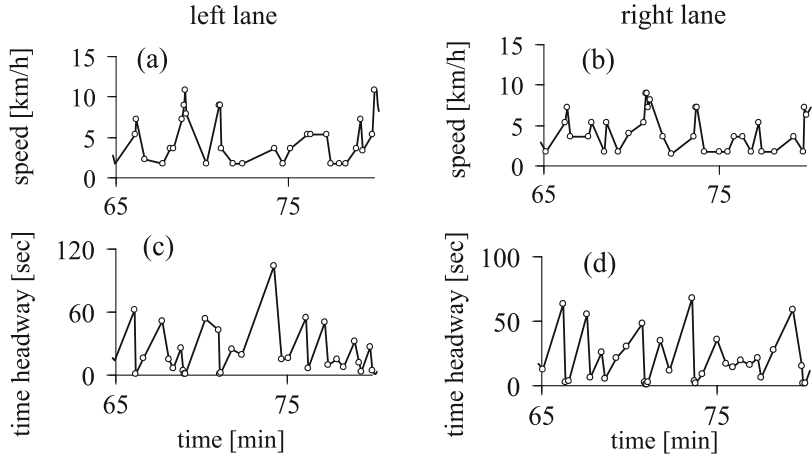
Microscopic Spatiotemporal Structure of Mega-jam

The crucial difference between GPs and a mega-jam, which occurs under condition (19), is as follows: Far enough upstream of the bottleneck, single-vehicle data of a GP shows a sequence of wide moving jams separated by non-interrupted flows (Fig. 69); in contrast, within the mega-jam, we can distinguish *no* sequence of wide moving jams (Fig. 73a, b). Rather than a sequence of wide moving jams, within the mega-jam, there is a complex sequence of upstream-moving flow interruption intervals (black regions in Fig. 74) for which criterion (23) is satisfied (Fig. 73c, d). These flow interruption intervals are separated by short time intervals within which vehicles move with low speeds (Fig. 73c, d). The latter time intervals are associated with moving blanks (white regions in Fig. 74).

Thus, as the microscopic spatiotemporal structure of a wide moving jam, the microscopic structure of a mega-jam consists of an alternation of flow interruption intervals and moving blanks (Fig. 74). This explain the term a *mega-jam* that is also a wide moving jam, however, with an extremely large width (in the longitudinal direction) continuously growing over time. The continuous growth of the mega-jam width occurs as long as the flow rate upstream of the mega-jam q_{in} exceeds $q^{(cong)}$ (Fig. 64d). As for GPs with a non-regular pinch region, within the mega-jam, we have found a very complex and non-regular spatiotemporal dynamics of flow interruption intervals and moving blanks. However, because within the mega-jam there are no wide moving jams separated by non-interrupted flows, the non-regular mega-jam dynamics is associated with the dynamic effects (i), (iii),

Spatiotemporal Features of Traffic Congestion,

Fig. 73 Simulated single-vehicle data for speed (a, b) and time headway (c, d) measured by a virtual detector at location $x = 5$ km related to $T_B = 60$ s. Left and right figures are related to the left and right lanes, respectively. The upstream boundary of the bottleneck is at location $x = 16$ km. (Adapted from Kerner 2009a)



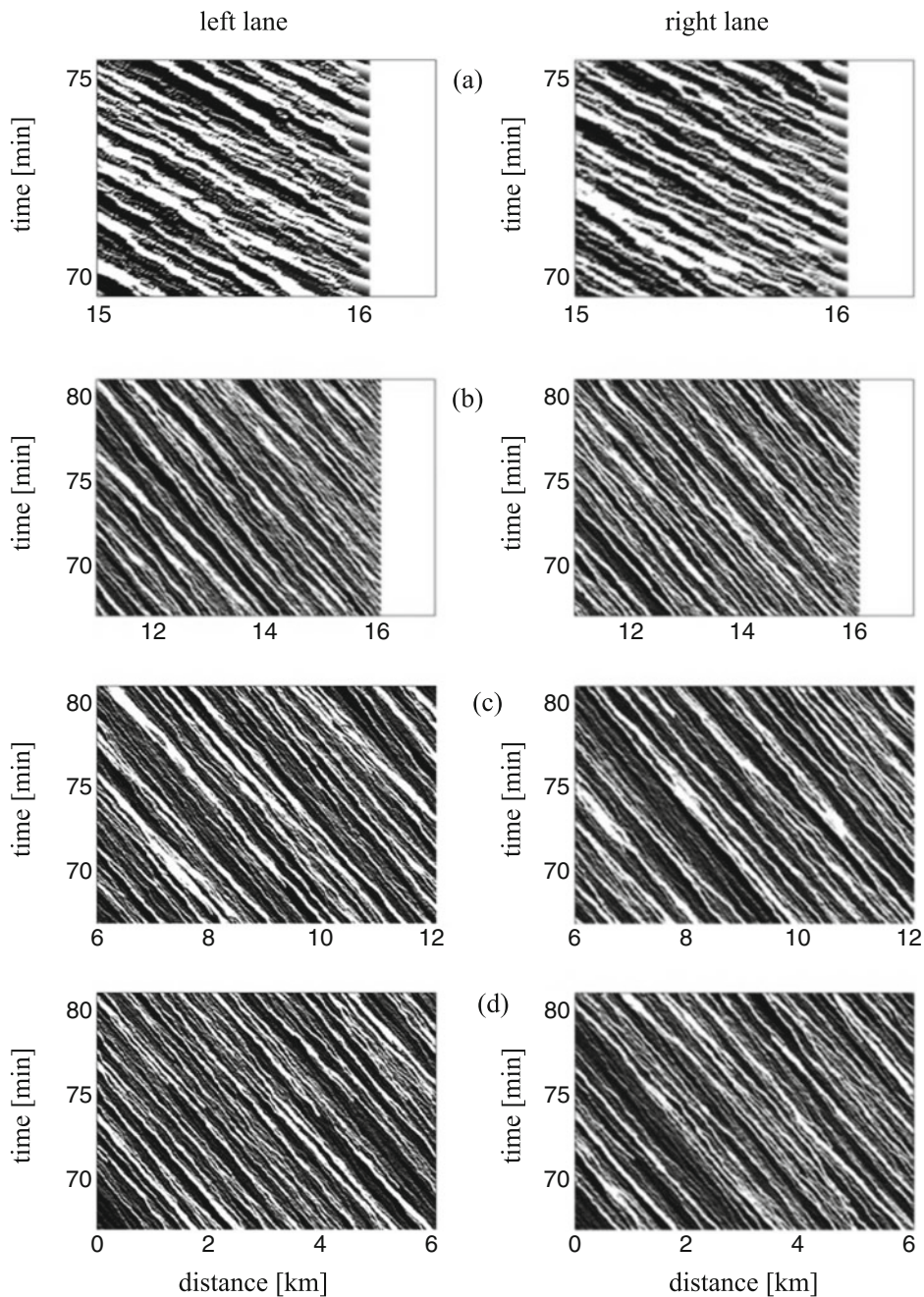
(v), and (vii) of section “[Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics](#)” only.

Physics of Non-regular Spatiotemporal Dynamics of Traffic Congestion at Heavy Bottlenecks

A phase transition from synchronized flow to a wide moving jam (S→J transition) occurs in a metastable synchronized flow state with respect to the S→J transition (section “[Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow](#)”). For this reason, the S→J transition is characterized by a random time delay T_{SJ} (Kerner 2004b, 2017a). T_{SJ} includes a random time delay of spontaneous nucleation of a narrow moving jam and a random time of the jam growth within the pinch region until the narrow moving jam transforms into a wide moving jam at the upstream boundary of the pinch region. Thus, the smaller the T_{SJ} , the shorter the $L^{(\text{pinch})}$. A random character of T_{SJ} explains a random character of the time dependence of $L^{(\text{pinch})}(t)$ (Fig. 67c, e, g). We have found that the larger the bottleneck strength, i.e., the smaller $q^{(\text{cong})}$, the lower the speed and the larger the density within the pinch region. However, the lower the speed and the larger the density in synchronized flow, the smaller the nucleus required for the emergence of a flow interruption interval, i.e., the smaller the mean random time delay $T_{SJ}^{(\text{mean})}$ of the S→J transition (Kerner 2009a, 2004b).

Thus, the larger the bottleneck strength, i.e., the smaller $q^{(\text{cong})}$, the larger the probability for the occurrence of a negligibly small T_{SJ} and the emergence of wide moving jams directly upstream of the bottleneck; in the latter case, $L^{(\text{pinch})} = 0$. This explains why under condition (17) the regular structure of the GP breaks due to the random disappearance of the pinch region during some time intervals. This explains also why $L_{\text{mean}}^{(\text{pinch})}$ decreases up to zero, when a strong increase in the bottleneck strength resulting in the subsequent strong decrease of $q^{(\text{cong})}$ causes a decrease in $T_{SJ}^{(\text{mean})}$ to a small value that is close to zero, i.e., when under condition (18), the pinch region does not exist.

The larger the bottleneck strength, the smaller the flow rate $q^{(\text{cong})}$ within a congested pattern. Therefore, the larger the bottleneck strength, the smaller the mean value of the synchronized flow speed and the larger the density within both the pinch region of synchronized flow of a GP and in the synchronized flow between wide moving jams (compare synchronized flow speeds between wide moving jams in Fig. 69 for different T_B). For this reason, the probability of the emergence of flow interruption intervals between wide moving jams, i.e., the emergence of new wide moving jams, increases rapidly with the increase in the bottleneck strength. As a result, many short flow interruption intervals appear at larger bottleneck strengths. On the other hand, short flow interruption intervals can dissolve easily, when the density decreases randomly upstream of these



Spatiotemporal Features of Traffic Congestion, Fig. 74 Fragments of Fig. 70d in larger scales in the left (figures left) and right lanes (right). White regions (single-vehicle speeds higher than 1.8 km/h) are related to moving

blanks. Black regions (single-vehicle speeds are equal to zero) are related to flow interruption intervals. $T_B = 60$ s. The upstream boundary of the bottleneck is at location $x = 16$ km. (Adapted from Kerner 2009a)

interruption intervals. A spatiotemporal competition between the random emergence and dissolution of flow interruption intervals can explain the

non-regular spatiotemporal jam dynamics of section “Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics.”

There are several sources for random speed disturbances in the traffic flow model used for simulations considered in sections “[Evolution of Traffic Phases at Heavy Bottlenecks](#)”–“[Microscopic Spatiotemporal Structure of Mega-jam](#),” whose growth leads to the emergence of flow interruption intervals in synchronized flows, for example, lane changing and the variety of the random delays in vehicle acceleration and decelerations (Kerner 2009a, 2004b). In addition, within the bottleneck, vehicles are forced to move at much longer safe time headway $\tau_{\text{safe}} = T_B$ than the safe time headway τ_{safe} outside the bottleneck. For this reason, the mean amplitude of speed disturbances in the pinch region is considerably larger than the one in synchronized flows between wide moving jams on a homogeneous road. This explains why the non-regular jam dynamics is considerably visible only at a bottleneck strength that is larger than the critical bottleneck strength for the occurrence of GPs with a non-regular pinch region.

When the bottleneck strength increases strongly and condition (19) is satisfied, wide moving jams merge into a mega-jam. The whole flow rate within any mega-jam is supplied by moving blanks *only*. As a result, under condition (16) we get

$$q^{(\text{cong})} = q_{\text{mega}}^{(\text{blanks})}, \quad (24)$$

where $q_{\text{mega}}^{(\text{blanks})}$ denotes the average flow rate associated with moving blanks within the mega-jam, which satisfies condition

$$q_{\text{mega}}^{(\text{blanks})} \leq q^{(\text{blanks})}, \quad (25)$$

where, as defined above, $q^{(\text{blanks})}$ is the average flow rate associated with moving blanks within wide moving jams separated by non-interrupted flows. The equality in (25) is related to (15). Thus, if under condition (16) the bottleneck strength increases and therefore $q^{(\text{cong})}$ decreases, then in accordance with (24), the flow rate $q_{\text{mega}}^{(\text{blanks})}$ decreases.

As mentioned in section “[Evolution of Traffic Phases at Heavy Bottlenecks](#)” (see (21)), the threshold bottleneck strength for the pinch region existence is larger than the critical bottleneck

strength for the mega-jam formation. To explain that this result has a general character, let us assume that the bottleneck strength is initially larger than the critical one for the mega-jam formation. In this case, $q^{(\text{cong})} = q_{\text{mega}}^{(\text{blanks})} < q^{(\text{blanks})}$. If now the bottleneck strength decreases gradually, then the flow rate $q^{(\text{cong})}$ increases. The flow rate $q^{(\text{cong})}$ can be supplied by moving blanks only up to the critical bottleneck strength satisfying condition (15), i.e., $q^{(\text{cong})} = q^{(\text{blanks})}$. When the bottleneck strength decreases further and becomes smaller than the critical one for the mega-jam formation, then $q^{(\text{cong})} > q^{(\text{blanks})}$. The difference $\Delta q_{\text{blanks}} = q^{(\text{cong})} - q^{(\text{blanks})}$ should be supplied by vehicles accelerating from wide moving jams; this means that wide moving jams separated by synchronized flows must appear in congested traffic. In particular, such synchronized flows appear just upstream of the bottleneck as a result of the upstream propagation of wide moving jams that emerge directly upstream of the bottleneck. Nevertheless, a pinch region of the synchronized flow at the bottleneck does not still appear as long as the difference Δq_{blanks} is small enough. Indeed, the pinch region appears only when at least one of the wide moving jams occurs at a *finite distance* upstream of the bottleneck in the synchronized flow due to a growing narrow moving jam that emerges initially in this flow. This is possible only if the former wide moving jam due to its upstream propagation is at a large enough distance upstream of the bottleneck. The latter can be realized when the difference Δq_{blanks} exceeds also some *finite value*, i.e., the threshold bottleneck strength for the pinch region existence should be always smaller than the critical one for the mega-jam formation. Thus, condition (21) is a general result for any heavy bottleneck.

Comparison with Empirical Results

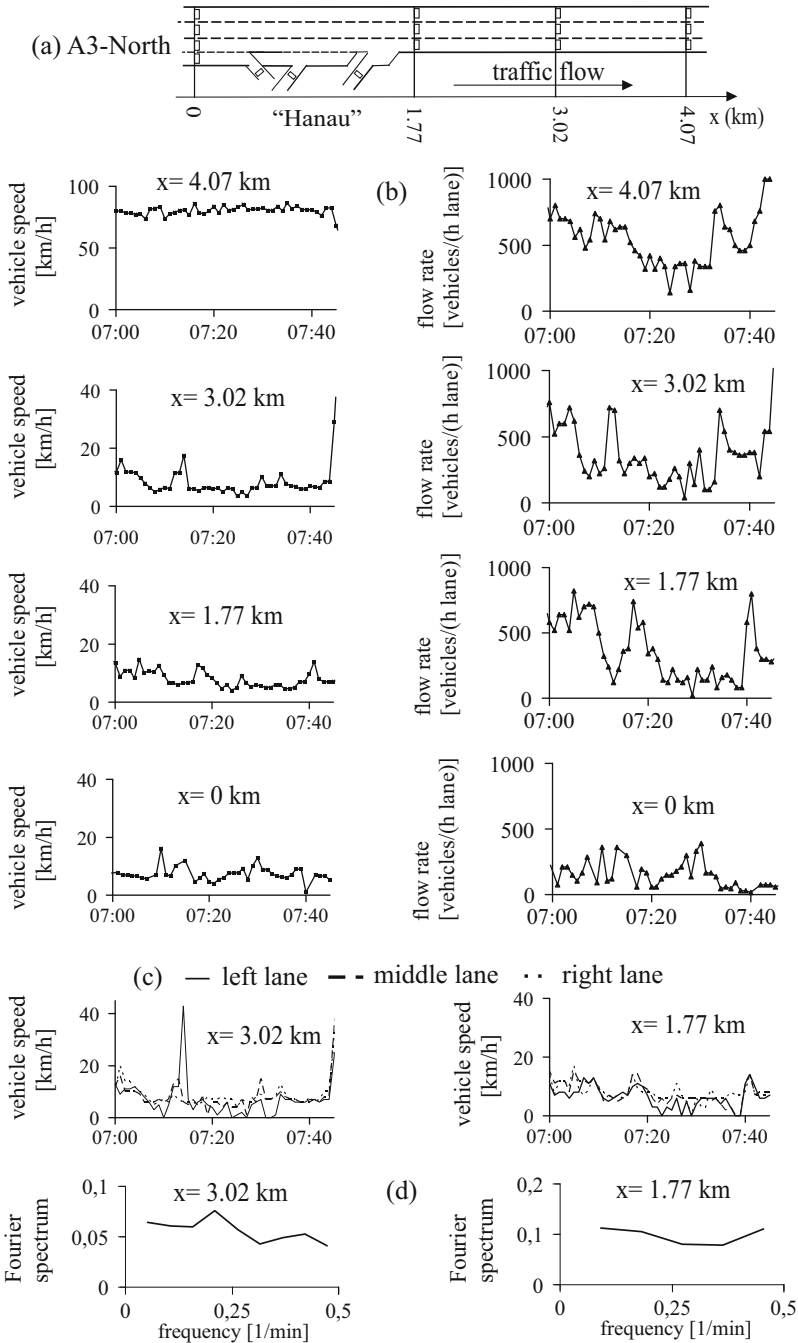
To compare the above theory of traffic congestion at heavy bottlenecks (Kerner 2007, 2008) with empirical congested patterns, one should have measured data for traffic congestion at a bottleneck, whose strength should be manually continuously changeable from the one associated with usual bottlenecks like on- and off-ramps to large

bottleneck strengths associated with heavy bottlenecks caused by bad weather conditions or accidents. Unfortunately, such measured data is not available. However, we can compare the theory with measured data related to two limiting cases:

- (i) Traffic congestion at a usual on-ramp bottleneck (Fig. 49)
- (ii) Traffic congestion at a heavy bottleneck caused by bad weather conditions – snow and ice on a road (Fig. 75)

Spatiotemporal Features of Traffic Congestion,

Fig. 75 Empirical structure of congestion caused by snow and ice: (a) Scheme of road detector arrangement on a section of the freeway A3-North in Germany near the intersection “Hanau.” (b) Average speed and flow rate across the freeway at different locations. (c) Speed in different lanes at two locations. (d) Fourier spectra of the speed in the left lane at two locations. 1-min average data. (Adapted from Kerner 2009a)



The speed distributions with congested patterns found in simulations (Fig. 64a, b) for the range of $T_B = 1.6 - 2.4$ s are associated with the flow rate $q^{(\text{cong})} = q^{(\text{pinch})}$ that is within the range

$$q^{(\text{cong})} = q^{(\text{pinch})} = 1120 - 1800 \text{ vehicles}/(\text{h lane}). \quad (26)$$

These speed distributions are qualitatively the same as those in an empirical GP shown in Fig. 49 and in all other known empirical GPs (Kerner 2004b). Moreover, values of empirical flow rates $q^{(\text{pinch})}$ (12) are approximately associated with the theoretical result (26). In particular, within the pinch region of the GP shown in Fig. 49a, the flow rate averaged between 7:00 and 8:00 and across the road is $q^{(\text{pinch})} = 1200$ vehicles/(h lane).

In contrast, for $T_B > 6$ s associated with the average flow rate $q^{(\text{cong})} < 625$ vehicles/(h lane), simulated congested patterns at heavy bottlenecks (Fig. 64c, d) are qualitatively the same as those we found in empirical data measured on many various days (and years) on different freeways, when very heavy bottlenecks caused by bad weather conditions or accidents occur for which empirical flow rates

$$q^{(\text{cong})} \lesssim 600 \text{ vehicles}/(\text{h lane}). \quad (27)$$

In this case, rather than regular structure of traffic congestion of GPs (Figs. 49 and 64a, b), both theoretical (Figs. 64c, d, 66i) and empirical traffic congested patterns exhibit non-regular spatiotemporal structure of congestion in which *no* sequences of wide moving jams can be distinguished in 1-min average data.

An example of such an empirical congested traffic pattern is shown in Fig. 75. Rather than the regular structure of congestion within the GP (Fig. 49), in measured data associated with bad weather conditions, a non-regular spatiotemporal structure of congestion is observed (Fig. 75). A heavy bottleneck appears on February 02, 2006, between locations 4.07 and 3.02 km due to snow and ice. Upstream of the bottleneck, very low speed and flow rate patterns ($x \leq 3.02$ km in

Fig. 75b) are observed. Downstream of the bottleneck ($x = 4.07$ km), vehicles have escaped from a congested pattern (speed is high); however, the flow rate is very small because the bottleneck reduced the average flow rate within the congestion strongly. For example, at $x = 3.02$ km, the flow rate $q^{(\text{cong})}$ averaged between 7:00 and 7:40 and across the road is 513 vehicles/(h lane). In contrast with the GP shown in Fig. 49, within traffic congestion shown in Fig. 75, non-regular low-speed patterns are observed in which *no* sequence of wide moving jams can be distinguished ($x \leq 3.02$ km). This conclusion is regardless of the flow rates to on- and off-ramps, which in the data set lead to an increase in $q^{(\text{cong})}$ at $x = 1.77$ km in comparison with $q^{(\text{cong})}$ at $x = 0$ km (Fig. 75a, b). If speeds in different lanes are compared (Fig. 75c), we find even more non-regular speed time dependencies: whereas no vehicles pass a detector in one of the lanes, i.e., the speed is zero, during the same time interval, the average speed in other lanes can be higher than zero. This explains why in time dependencies of speeds averaged across the road (Fig. 75b), this very low speed is seldom equal to exactly zero.

As congested traffic in measured data (Fig. 75b, c), simulated congested traffic at a heavy bottleneck is also nonhomogeneous in space and time (Fig. 66i, j). In the measured and simulated data, as followed from Fourier spectra of speed time dependencies, these non-homogeneous congested patterns exhibit non-regular spatiotemporal behavior (Figs. 66i and 75d). This is in contrast with the regular structure of traffic congestion within the GP (Fig. 49).

As in measured data (Fig. 75b, c), in 1-min average data related to a traffic congested pattern found in simulations (Fig. 64c), very low speed and flow rate are realized upstream of the bottleneck. *No* sequence of wide moving jams can be distinguished in the congested pattern shown in Figs. 64c and 66i. Downstream of the bottleneck ($x > 16.3$ km), vehicles have escaped from the congested pattern (speed is high) (Fig. 64c); however, the flow rate is very small because the bottleneck reduced the average flow rate within the congested pattern strongly.

A summary of the above comparison of theory of traffic congestion with measured data is as follows:

- (i) Spatiotemporal distributions of the average speed and flow rate within GPs occurring at usual bottlenecks are qualitatively the same in the above theory and measured data (see sections “[Congested Patterns at Road Bottlenecks](#)” and “[Microscopic Features of Traffic Phases in Congested Traffic](#)”): in both simulated and measured data, GPs consist of the pinch region and wide moving jams upstream. Measured and simulated speed time dependencies show *regular* character of wide moving jam propagation.
- (ii) In contrast with item (i), in 1-min average data related to simulated congested patterns and to measured data for traffic congestion at heavy bottlenecks, *no* sequence of wide moving jams can be distinguished. In simulations and measured data, spatiotemporal distributions of the average speed and flow rate in these congested patterns are nonhomogeneous in space and time. Fourier spectra of the associated nonhomogeneous time dependencies of the speed show non-regular character of traffic congestion at the heavy bottlenecks both in simulations (Fig. 66l) and measured data (Fig. 75d).

As followed from sections “[Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics](#)” and “[Microscopic Spatiotemporal Structure of Mega-jam](#),” the non-regular pattern behavior of item (ii) can be explained by two different reasons:

1. The occurrence of a GP with the non-regular dynamics of wide moving jams
2. The occurrence of a mega-jam

However, based on 1-min average data, these two reasons of non-regular traffic congestion cannot be distinguished from each other. Indeed, theoretical results of sections “[Microscopic Spatiotemporal Features of Non-regular Moving Jam](#)

[Dynamics](#)” and “[Microscopic Spatiotemporal Structure of Mega-jam](#)” allow us to assume that at a very heavy bottleneck caused, for example, by bad weather conditions or accidents, there should be a very fine microscopic spatiotemporal structure of traffic congestion associated with non-regular spatiotemporal dynamics of wide moving jams or/and flow interruption intervals as well as moving blanks. However, this theoretical fine structure of traffic congestion, even if it exists in real traffic, would be averaged in 1-min average measured data, i.e., they could not be found in this data. Thus, to make a more detailed comparison of the theory of traffic congestion at heavy bottlenecks with measured data, single-vehicle data measured within traffic congestion at heavy bottlenecks is required. Unfortunately, a study of empirical single-vehicle data for traffic congested at heavy bottlenecks has not been made up to now. Such a comparison of the theory of traffic congestion at heavy bottlenecks presented in section “[Congested Patterns at Heavy Bottlenecks](#)” with measured single-vehicle data is a separate and important task of further investigations.

“Jam-Absorption” Effect in Highway and City Traffic

Up to now we have considered congested traffic in which, for example, due to a large on-ramp inflow rate at an on-ramp bottleneck (section “[Heavy Traffic Congestion](#)”) or due to the existence of a heavy bottleneck (section “[Congested Patterns at Heavy Bottlenecks](#)”), moving jams emerge spontaneously in synchronized flow. However, in Kerner (2012), we have shown that there is a special driver behavior called “strong” driver speed adaptation at which all moving jams dissolve in congested traffic. As a result, congested traffic consists of the synchronized flow traffic phase in which no moving jams emerge. Following Kerner (2012), in this section, we discuss briefly such an effect of moving jam dissolution in congested traffic that is also called “jam-absorption” effect.

Driver's Choice of Space Gap in Synchronized Flow: Strong and Weak Speed Adaptation

In accordance with a hypothesis of the three-phase theory about a 2D region of synchronized flow states (Kerner 1998c, 1999a), at a given vehicle speed v in synchronized flow, there are the *infinite number* of space gaps g within the range

$$g_{\text{safe}} \leq g \leq G \quad (28)$$

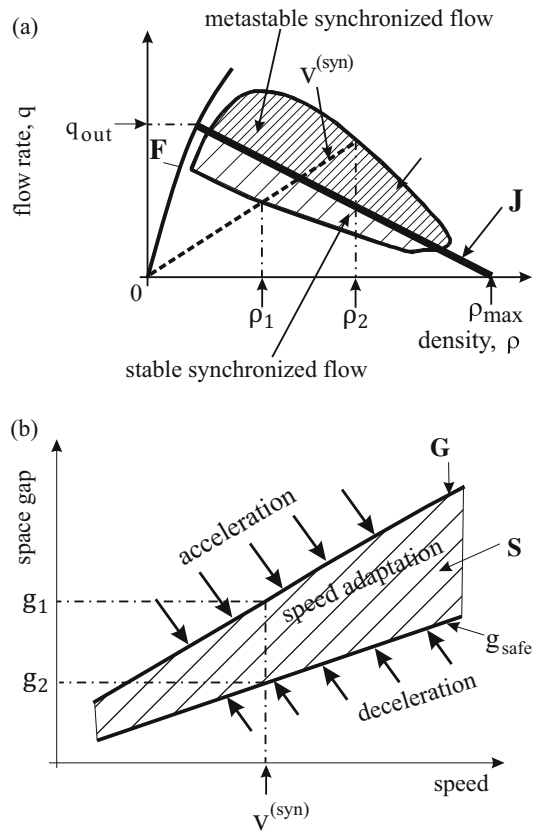
at which a driver can move with this speed v . In (28), g is a space gap between the vehicle and the preceding vehicle, G is a synchronization space gap, and g_{safe} is a safe space gap (see explanations of G and g_{safe} in the book (Kerner 2017a) or in the review article (Kerner 2018a) of this Encyclopedia).

When a driver approaches a slower preceding vehicle moving at a time-independent speed $v^{(\text{syn})}$ (Fig. 76a) and the driver cannot pass it, the driver decelerates within the synchronization space gap G adapting the speed to the speed of the preceding vehicle $v^{(\text{syn})}$ (Fig. 76a, b). This speed adaptation within the synchronization gap is associated with the 2D region of states of synchronized flow given by conditions (28) (Fig. 76b). The speed adaptation effect (labeled by “speed adaptation” in Fig. 76b) is the adaptation of the vehicle speed to the speed of the preceding vehicle at any space gap within the space gap range (28), i.e., without caring what the precise space gap to the preceding vehicle is.

Thus, in the three-phase theory, while moving in synchronized flow, a driver can arbitrarily choose a space gap to the preceding vehicle within the 2D region of synchronized flow (Fig. 76) (Kerner 1998b, c, 1999a, 2004b). In accordance with Kerner (2012), we distinguish two limit cases of driver speed adaptation within the 2D region of synchronized flow (Figs. 76 and 77):

- (i) “Strong” driver speed adaptation (Fig. 77a)
- (ii) “Weak” driver speed adaptation (Fig. 77b)

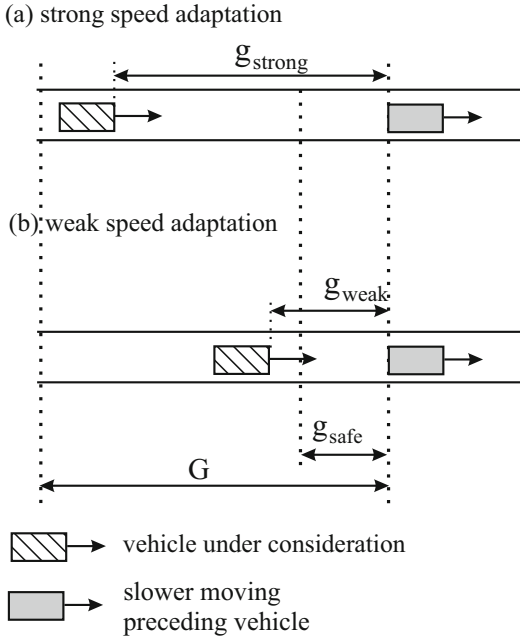
Under strong speed adaptation, a driver chooses a large mean speed gap g_{strong} (Fig. 77a) that satisfies conditions



Spatiotemporal Features of Traffic Congestion, Fig. 76 Qualitative explanation of car-following in the three-phase theory (Kerner 1998c, a, 1999a, 2004b, 2009c): (a) Free flow (F), 2D region of synchronized flow states, and line J in the flow–density plane taken from Fig. 22a. (b) A part of the 2D region of synchronized flow states from (a) presented in the speed–space-gap plane. In (b), a vehicle accelerates at $g > G$ (labeled by “acceleration”) and decelerates at $g < g_{\text{safe}}$ (labeled by “deceleration”), whereas under condition (28), the vehicle adapts its speed to the speed of the preceding vehicle without caring what the precise space gap is (labeled by “speed adaptation”); S – synchronized flow. $v^{(\text{syn})}$ is a constant speed in synchronized flow. (Adapted from Kerner 2004b)

$$g_{\text{strong}} \leq G, \quad G - g_{\text{strong}} \ll g_{\text{strong}} - g_{\text{safe}}. \quad (29)$$

In contrast, under weak speed adaptation, a driver chooses a small mean speed gap g_{weak} (Fig. 77b) that satisfies conditions



Spatiotemporal Features of Traffic Congestion, Fig. 77 Qualitative explanation of strong driver speed adaptation (a) and weak driver speed adaptation (b) within 2D region of states of synchronized flow shown by dashed region in Fig. 76b in which the speed adaptation effect is realized. (Adapted from Kerner 2012)

$$g_{\text{weak}} \geq g_{\text{safe}}, \quad g_{\text{weak}} - g_{\text{safe}} \ll G - g_{\text{weak}}. \quad (30)$$

Stochastic simulations of the driver's choice of a space gap within the 2D region of synchronized flow have been discussed in Appendix A of the book (Kerner 2017a).

Formulas (29) and (30) can be easily understood, if we discuss the relation of the speed dependencies of the synchronization space gap $G(v)$ and the safe space gap $g_{\text{safe}}(v)$ shown in Fig. 76b to the 2D states of synchronized flow in the flow–density plane shown in Fig. 76a. We consider some states of synchronized flow related to the same constant speed $v = v^{(\text{syn})}$. We can see that for the space gaps g_1 and g_2 shown in Fig. 76b, we get

$$g_1 = G(v^{(\text{syn})}), \quad g_2 = g_{\text{safe}}(v^{(\text{syn})}). \quad (31)$$

These space gaps g_1 and g_2 determine, respectively, the vehicle densities $\rho = \rho_1$ and $\rho = \rho_2$ given by formulas

$$\rho_1 = \frac{1}{G(v^{(\text{syn})}) + d} \quad \text{and} \quad \rho_2 = \frac{1}{g_{\text{safe}}(v^{(\text{syn})}) + d} \quad (32)$$

at which the line of the constant speed $v^{(\text{syn})}$ intersects in the flow–density plane, respectively, the upper boundary and the low boundary of the 2D states of synchronized flow (Fig. 76a). In (32), d is the vehicle length that includes the mean space gap between vehicles that are in a stop within a wide moving jam.

This consideration explains that at weak driver speed adaptation (30) synchronized flow states are related to a 2D region in a vicinity of the upper boundary of the 2D states of synchronized flow. The states of synchronized flow are metastable with respect to an $S \rightarrow J$ transition (“metastable synchronized flow” in Fig. 76a). Contrarily, at strong driver speed adaptation (29) synchronized flow states are related to a 2D region in a vicinity of the low boundary of the 2D states of synchronized flow. These states of synchronized flow are stable with respect to an $S \rightarrow J$ transition (“stable synchronized flow” in Fig. 76a).

Transformation of General Pattern into Widening Synchronized Flow Pattern

As explained in section “Hypotheses About Nucleation Nature of Moving Jam Emergence in Synchronized Flow,” accordingly to the three-phase theory (Kerner 1998c), synchronized flow states that are on and above the line J in the flow–density plane are metastable states with respect to an $S \rightarrow J$ transition. Contrarily, synchronized flow states that are below line J in the flow–density plane are stable states with respect to an $S \rightarrow J$ transition (see Fig. 22). As found in Kerner (2012), the probability of occurrence of either metastable synchronized flow states or stable synchronized flow states in congested traffic depends basically on the characteristics of driver speed adaptation in synchronized flow.

Under weak speed adaptation (Fig. 77b), on average, drivers come relatively closely to associated preceding vehicles. As a result, the average time headway between drivers is relatively small in synchronized flow. This results in states of synchronized flow that are usually above the line

J in the flow–density plane. In these metastable states of synchronized flow (“metastable synchronized flow” in Fig. 76a), speed disturbances of relatively small amplitudes are nuclei for the $S \rightarrow J$ instability. Therefore, narrow moving jams can easily emerge spontaneously in the synchronized flow (see section “Moving Jam Emergence in Framework of Three-Phase Traffic Theory”). In this case, when the bottleneck strength is large enough (e.g., the on-ramp inflow rate is large enough at an on-ramp bottleneck), a general pattern (GP) can emerge at a highway bottleneck (see section “General Congested Patterns”). Congested traffic of the GP consists of the synchronized flow and wide moving jam phases (Fig. 78a).

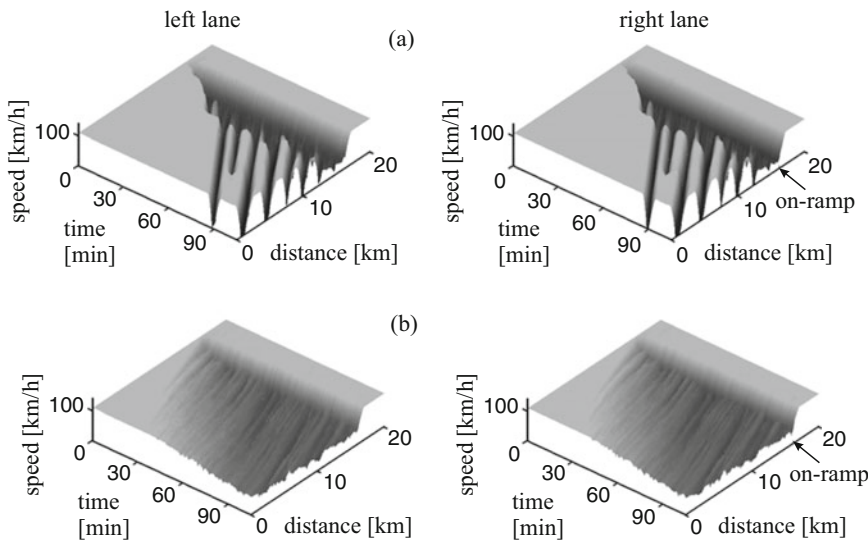
In the opposite case of strong driver speed adaptation (Fig. 77a), a driver that approaches a slower-moving preceding vehicle adapts the speed to that of the preceding vehicle quickly. In this case, while a vehicle decelerates, its space gap does not decrease considerably. This results in states of synchronized flow that are below the line J (“stable synchronized flow” in Fig. 76a). In these stable states of synchronized flow with respect to moving jam emergence, no moving

jams can emerge or persist over time. Therefore, all moving jams dissolve in the synchronized flow. As a result, rather than the GP (Fig. 78a), at the same other model parameters, a WSP remains at the bottleneck in which no moving jams can emerge (Fig. 78b) (Kerner 2012). Such an effect of moving jam dissolution can be called the “jam-absorption” effect or the effect of moving jam dissolution in synchronized flow:

- The jam-absorption effect in synchronized flow occurs, when a large enough mean space gap (mean time headway) between vehicles in synchronized flow upstream of a moving jam is realized that is related to a point in the flow–density plane lying below the line J (“stable synchronized flow” in Fig. 76a). In this case, the moving jam dissolves in the synchronized flow (Kerner 2012).

Diagram of Congested Patterns Under Strong Speed Adaptation

We have already mentioned that the main effect of strong speed adaptation within 2D region of synchronized flow states is the suppression of the



Spatiotemporal Features of Traffic Congestion, Fig. 78 Simulations of transformation of GP (a) into widened synchronized flow pattern (WSP) (b) through driver’s choice of stronger speed adaptation parameters: in (b), the driver speed adaptation is considerably stronger than in (a), while the flow rate on the main road upstream

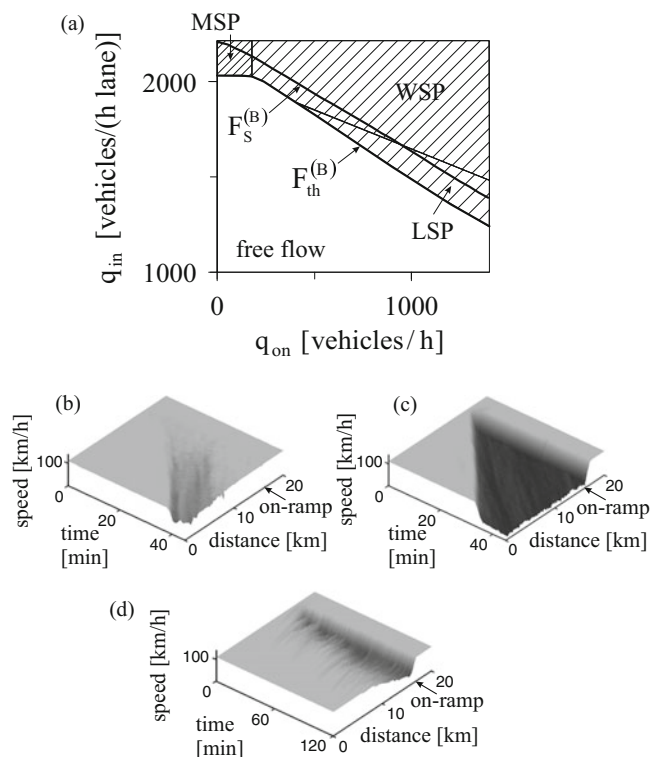
of the bottleneck and the on-ramp inflow as well as other model parameters are the same in (a) and (b). Simulations have been made on a two-lane road with an on-ramp bottleneck. Vehicle speed in space and time. Figures in left panel, left lane; right panel, right lane. (Adapted from Fig. 14 of Kerner 2012)

emergence of moving jams in synchronized flow. This can be seen when we compare the diagram of congested patterns at an on-ramp bottleneck under a relatively weak speed adaptation (Fig. 56a) with the diagram of congested patterns under strong speed adaptation (Fig. 79a).

At any point (q_{on}, q_{in}) related to the boundary $F_S^{(B)}$ in the diagram, traffic breakdown ($F \rightarrow S$ transition) occurs at the bottleneck with probability $P^{(B)} = 1$ during a given time interval T_{ob} for observing free flow at the bottleneck ($T_{ob} = 30$ min in Fig. 79a). In other words, the boundary $F_S^{(B)}$ determines the dependence of the maximum highway capacity $C_{max}(q_{in}, q_{on})$ on the flow rates q_{on} and q_{in} . As in Fig. 56a, at the boundary $F_S^{(B)}$, different SPs emerge spontaneously (Fig. 79a). Contrary to the diagram shown in Fig. 56a, when the on-ramp inflow rate q_{on} increases, subsequently *no* wide moving jams emerge spontaneously in synchronized flow of the SP (Fig. 79a). This means that GPs do not appear even at low speeds and large vehicle densities within synchronized flow of the SP (Fig. 79a).

Spatiotemporal Features of Traffic Congestion,

Fig. 79 Simulated synchronized flow patterns (SP) at on-ramp bottleneck under strong speed adaptation: (a) Diagram of congested patterns at on-ramp bottleneck. (b–d) Speed in space and time in the left lane within moving SP (MSP) (b), widening SP (WSP) (c), and localized SP (LSP) (d). (Adapted from Kerner 2012)



The boundary $F_{th}^{(B)}$ (Fig. 79a) is related to a threshold flow rate for traffic breakdown $q_{th}^{(B)}$ at which the probability $P^{(B)}$ that traffic breakdown occurs at the bottleneck during the time interval T_{ob} reaches its minimum value: left and below of the boundary $F_{th}^{(B)}$ the probability of traffic breakdown at the bottleneck during the time interval T_{ob} is equal to zero – $P^{(B)} = 0$. In other words, the boundary $F_{th}^{(B)}$ determines the dependence of the threshold flow rate for traffic breakdown $q_{th}^{(B)}(q_{in}, q_{on})$ on the flow rates q_{on} and q_{in} .

Dissolution of Moving Queues and Formation of Synchronized Flow in Oversaturated City Traffic

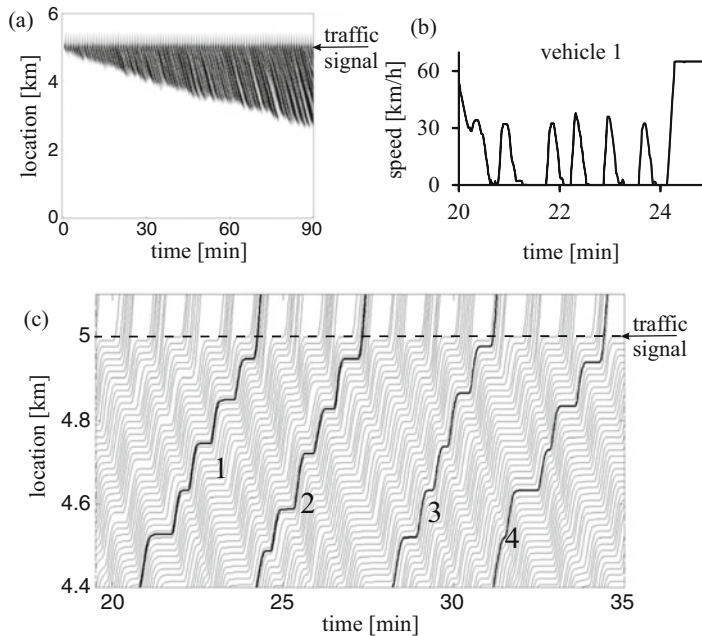
As mentioned in section “Introduction,” this article is devoted to a consideration of spatiotemporal congested patterns in highway traffic. This is because empirical features of traffic congestion in city traffic have not satisfactorily been studied. The situation has been changed only recently through measurements of empirical single-vehicle data. The recent studies of the empirical single-vehicle data in city traffic made in Kerner et al.

(2014a), Kaufmann et al. (2017, 2018) have confirmed the existence of SPs and, in particular, MSPs in city traffic occurring at traffic signals predicted theoretically in Kerner (2011, 2013a, 2014), Kerner et al. (2014b, 2013a).

It is well-known that there are two main types of city traffic at traffic signals: under- and oversaturated traffic. In undersaturated traffic, all vehicles, which are waiting within a queue during the red signal phase, can pass the signal during the green signal phase. An opposite case occurs in oversaturated traffic: some of the vehicles in the queue cannot pass the signal location during the green phase resulting in the queue growth (Fig. 80a, b) (Webster 1958) (for reviews, see, e.g., Gartner and Stamatiadis 2009; Newell 1982; Dion et al. 2004). In accordance with the

classical theory (Webster 1958; Gartner and Stamatiadis 2009; Newell 1982; Dion et al. 2004; Geroliminis and Skabardonis 2011), well-developed oversaturated traffic consists of a sequence of moving queues with stopped vehicles separated by regions in which vehicles move from one moving queue to the adjacent downstream-moving queue (Fig. 80c).

In the classical theory of city traffic at traffic signal, oversaturated city traffic should consist of an alternation of moving queues and free flow *only*. Contrary to the classical theory (Fig. 80a, b) (Webster 1958; Gartner and Stamatiadis 2009; Newell 1982; Dion et al. 2004; Geroliminis and Skabardonis 2011), in Kerner et al. (2013a), we have found that in oversaturated city traffic under strong speed adaptation, time headway between



Spatiotemporal Features of Traffic Congestion, Fig. 80 Simulated spatiotemporal structure of oversaturated traffic at traffic signal in the framework of the classical theory of city traffic: (a, b) Speed in space and time presented by regions with variable shades of gray (a) (in white regions the speed is higher than 50 km/h, in black regions the speed is zero) and microscopic (single-vehicle) speed along vehicle trajectory 1 (b) shown in (c). (c) Vehicle trajectories for some time interval and for 600 m long road section upstream of the signal location at $x = 5$ km from the beginning ($x = 0$) of a single-lane

road (each 2nd trajectory is shown). Signal parameters: $\vartheta = 60$ s, $T_R = 28$ s, $T_Y = 2$ s (ϑ , T_R , and T_Y are the durations of the signal cycle, the red signal phase, and the yellow signal phase, respectively). The arrival flow rate is time-independent: $q_{in} = 1000$ vehicles/h. Some chosen vehicle trajectories 1–4 in (c) illustrate vehicle behavior in oversaturated city traffic: in accordance with the classical theory (Webster 1958; Gartner and Stamatiadis 2009; Newell 1982; Dion et al. 2004; Geroliminis and Skabardonis 2011), oversaturated traffic consists of an alternation of moving queues and free flow *only*

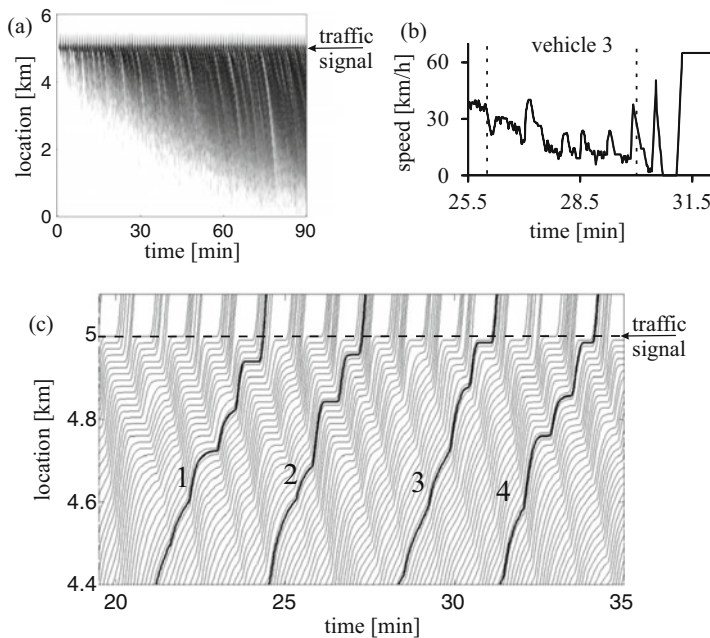
drivers increase and moving queues dissolve at some distance of traffic signal. As a result, moving synchronized flow patterns (MSPs) appear in city traffic (Fig. 81). We have found that the mean duration of a vehicle stop within a moving queue decreases, while the moving queue moves away upstream from the signal. At some distance of the signal location, moving queues can dissolve, and oversaturated traffic consists of synchronized flow only in which vehicles do not come to a stop (Fig. 81b, c). At strong enough speed adaptation, vehicles stop only a few (one or two) times in a neighborhood of the signal location (Fig. 81b, c). This means that all moving queues dissolve and transform into synchronized flow already at relatively short distances upstream of the signal location. In this case, in contrast with the classical theory (Fig. 80b, c), a vehicle accelerates and

decelerates many times during its motion in synchronized flow.

Empirical MSPs in oversaturated city traffic predicted in Kerner et al. (2013a) have been observed in Kerner et al. (2014a), Kaufmann et al. (2017, 2018) (Figs. 82 and 83).

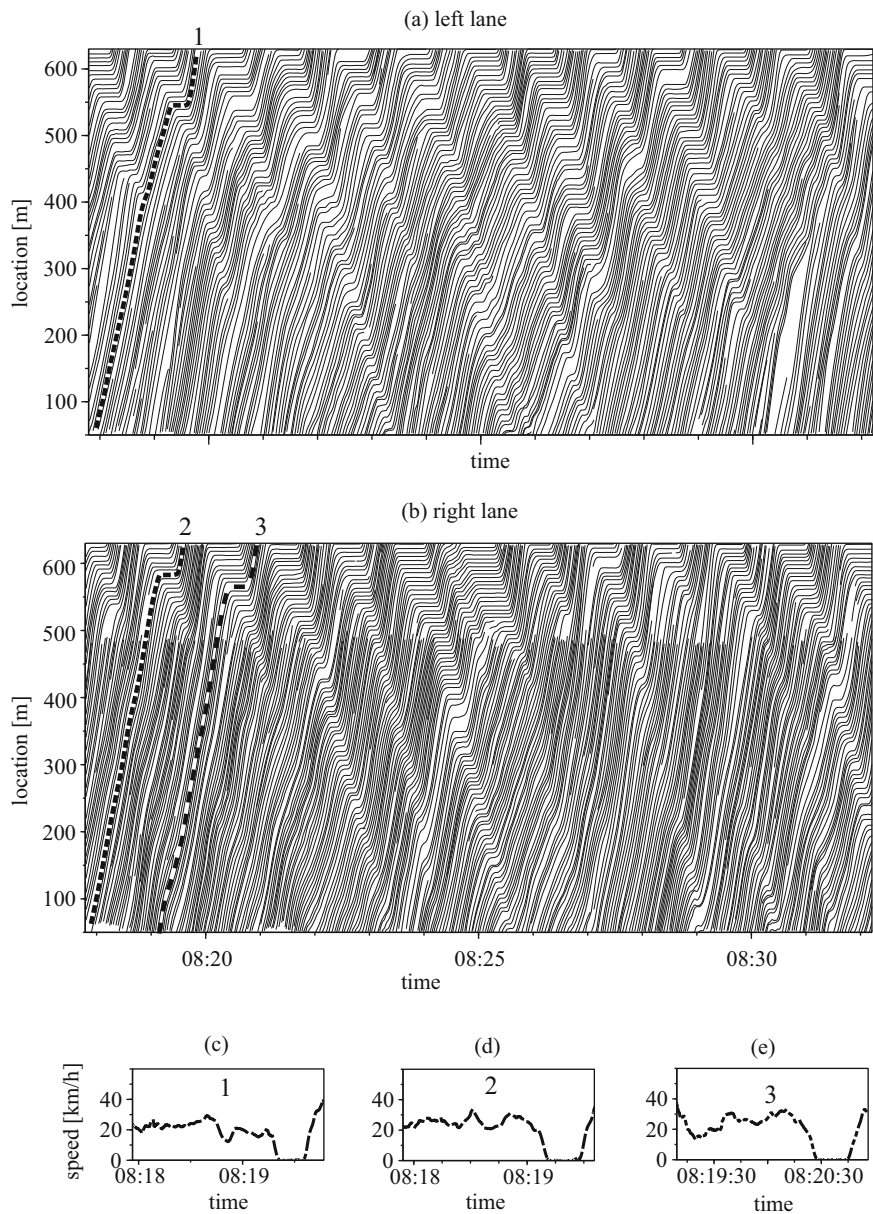
About Applications of Intelligent Transportation Systems in Congested Traffic

Under a large total traffic network inflow rate, it can turn out that congested traffic cannot be avoided in a traffic network. Due to heavy traffic congestion, travel time and fuel consumption as well as other travel costs can increase considerably in the network. For this reason, users of traffic and transportation networks can expect that even if traffic congestion cannot be avoided in traffic networks, nevertheless, through the



Spatiotemporal Features of Traffic Congestion, Fig. 81 Simulations of spatiotemporal effect of moving queue dissolution with occurrence of moving synchronized flow patterns (MSP) under strong speed adaptation: (a, b) Speed in space and time presented by regions with variable shades of gray (a) and microscopic speed along vehicle trajectory 3 (b) shown in (c). (c) Fragment of vehicle trajectories upstream of the signal (each 2nd trajectory is shown); some chosen vehicle trajectories 1–4

illustrate vehicle behavior in oversaturated city traffic under strong speed adaptation of the three-phase theory: rather than alternation of moving queues and free flow of the classical theory, we have found that at some distance upstream of the signal location, moving queues can dissolve and oversaturated traffic consists of synchronized flow only in which vehicles do not come to a stop. (Adapted from Kerner et al. 2013a)

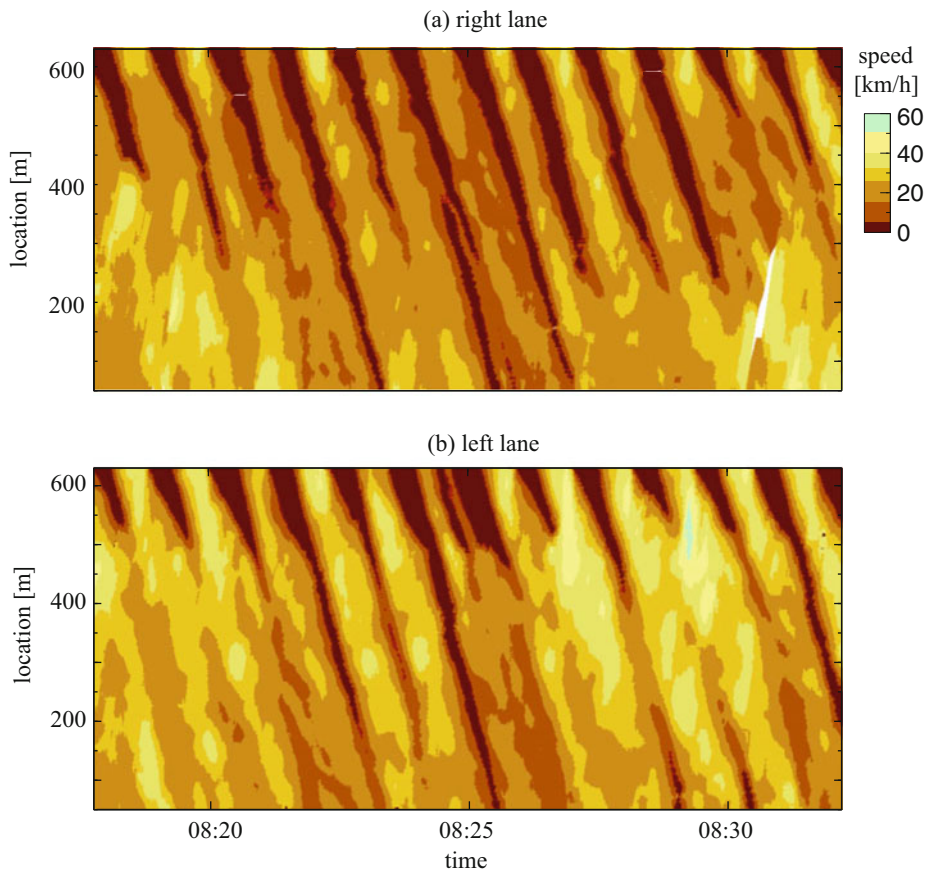


Spatiotemporal Features of Traffic Congestion, Fig. 82 Empirical observations of dissolution of moving queues at traffic signal and formation of moving synchronized flow patterns (MSP) in oversaturated city traffic at some distance upstream of the signal. Microscopic spatiotemporal measurements made on December 10, 2015:

complete vehicle trajectories in right lane (a), left lane (b), and microscopic (single-vehicle) speeds of three selected vehicles (c–e) according to vehicle trajectory numbers in (a, b). The position of the stop lane at the traffic signal is 640 m. (Adapted from Kaufmann et al. 2017, 2018)

application of intelligent transportation systems (ITS), fuel consumption can be decreased, whereas traffic safety and travel comfort can be increased in congested traffic.

In particular, in Hemmerle et al. (2016a, b), and Hermanns et al. (2015), it has been shown that vehicle fuel consumption in the synchronized flow traffic phase can be considerably smaller



Spatiotemporal Features of Traffic Congestion, Fig. 83 Continuation of Fig. 82. Empirical observations of dissolution of moving queues at traffic signal and formation of moving synchronized flow patterns (MSP)

in upstream of the signal oversaturated city traffic: the same speed data as those in Fig. 82 presented by regions with variable color. (Adapted from Kaufmann et al. 2017, 2018)

than that in congested traffic consisting of synchronized flow and wide moving jams. This explains the importance of the jam-absorption effect due to strong speed adaptation of the three-phase theory discussed in this section: if through the use of intelligent transportation systems (ITS) conditions of strong driver speed adaptation are realized in congested traffic, all moving jams dissolve in congested traffic. This will decrease fuel consumption in a traffic network together with the increase in traffic safety and travel comfort.

Conclusions: Fundamental Empirical Features of Spatiotemporal Congested Traffic Patterns

There are empirical features of phase transitions and spatiotemporal congested patterns at road bottlenecks that are reproducible in traffic observations on numerous days and years on different highways in various countries. Thus, these qualitative empirical features remain in traffic flow with very different driver behavioral characteristics and vehicle parameters. The fundamental

empirical spatiotemporal features of the phase transitions and congested traffic can be explained within the framework of the three-phase theory.

Empirical features of congested patterns at highway bottlenecks are as follows:

- (i) In congested traffic, two different traffic phases can be distinguished: the synchronized flow phase and the wide moving jam phase. These traffic phases are defined through the use of the spatiotemporal empirical (objective) criteria [J] and [S]. Thus, there are three traffic phases in traffic flow: (1) free flow, (2) synchronized flow, and (3) wide moving jam.
- (ii) Traffic breakdown in free flow at a highway bottleneck is associated with a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition). The $F \rightarrow S$ transition exhibits the nucleation nature. For this reason, traffic breakdown in free flow at the bottleneck can be either spontaneous or induced.
- (iii) Wide moving jams do not emerge spontaneously in free flow. Wide moving jams can emerge spontaneously only in synchronized flow; wide moving jams emerge due to the sequence of $F \rightarrow S \rightarrow J$ transitions.
- (iv) The larger the density in synchronized flow, the larger the frequency of spontaneous moving jam emergence in that synchronized flow. In synchronized flow of higher speeds and lower densities, wide moving jams should not necessarily emerge spontaneously.
- (v) A usual scenario of an $S \rightarrow J$ transition, i.e., the emergence of a wide moving jam in synchronized flow is as follows. Firstly, due to an $F \rightarrow S$ transition at a bottleneck, synchronized flow occurs at the bottleneck. Then, a compression of synchronized flow (pinch effect) is realized in this synchronized flow. In the pinch region of synchronized flow, an $S \rightarrow J$ instability occurs. Due to the $S \rightarrow J$ instability, narrow moving jams emerge spontaneously. The amplitude of a narrow moving jams can grow while the jam propagates upstream within the pinch region. The growing narrow moving jam can be considered a nucleus for the $S \rightarrow J$ transition: if a narrow moving jam grows continuously over time, it transforms into a wide moving jam ($S \rightarrow J$ transition).
- (vi) The development of a narrow moving jam ($S \rightarrow J$ instability) should not necessarily lead to the emergence of a wide moving jam ($S \rightarrow J$ transition). In a general case, the $S \rightarrow J$ instability should be strictly distinguished from the $S \rightarrow J$ transition.
- (vii) A narrow moving jam belongs to the synchronized flow phase. Contrarily, a wide moving jam belongs to the wide moving jam phase. Therefore, the narrow moving jam should be strictly distinguished from the wide moving jam.
- (viii) During the dynamics of the transformation of a narrow moving jam into a wide moving jam, the mean velocity of the downstream jam front tends to a characteristic velocity. Wide moving jams possess *characteristic* parameters and features that do not depend on initial conditions and disturbances in traffic flow. The characteristic parameters are the same for different wide moving jams. The characteristic parameters can depend on control parameters of traffic (weather, road conditions, etc.). These characteristic parameters and features are as follows:
 - (a) The mean velocity of the downstream front of a wide moving jam. The wide moving jam propagates upstream through any traffic state and through any bottleneck while maintaining the mean velocity of the downstream jam front.
 - (b) When free flow occurs in the wide moving jam outflow, the flow rate and density in the jam outflow are also characteristic parameters.
 - (c) The flow rate in the jam outflow is lower than the maximum possible flow rate in free flow.

- (ix) There are two main types of congested patterns at an isolated highway bottleneck: synchronized flow (SP) and general patterns (GP).
- (x) In contrast to wide moving jam propagation, an SP that propagates upstream can be caught at a bottleneck (catch effect) when the SP reaches the bottleneck. Due to the catch effect, an induced $F \rightarrow S$ transition (induced traffic breakdown) occurs at the bottleneck.
- (xi) There are three types of SPs: a moving SP (MSP), a widening SP (WSP), and a localized SP (LSP).
- (xii) There can be complex spontaneous transformations between various congested patterns at the bottleneck over time.
- (xiii) At the same traffic demand, various congested patterns can be formed at the bottleneck (the probabilistic nature of spatiotemporal congested patterns at the bottleneck).
- (xiv) If two or more adjacent effectual bottlenecks are close to one another, expanded congested patterns (EPs) often emerge. In an EP, synchronized flow affects at least two adjacent effectual bottlenecks.
- (xv) When a wide moving jam that has initially emerged downstream of a bottleneck propagates through synchronized flow upstream of the bottleneck, this foreign wide moving jam can considerably influence on other moving jams, which are just emerging in that synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam.
- (xvi) For each effectual highway bottleneck, or each set of several adjacent effectual bottlenecks at which congested patterns occur, the spatiotemporal structure of a congested pattern possesses some predictability, i.e., characteristic, unique, and reproducible features.
- (xvii) Rather than regular wide moving jam propagation observed upstream of usual bottlenecks like on- and off-ramp bottlenecks, no sequence of wide moving jams can be distinguished in 1-min average measured data for non-regular and non-homogeneous congested traffic, which is usually observed upstream of a heavy bottleneck. This empirical phenomenon can be explained by random disappearance and appearance of the pinch region of synchronized flow of a GP over time, by a non-regular dynamics of wide moving jams within the GP as well as by the merger of wide moving jams into a mega-wide moving jam (mega-jam) that should occur upstream of a very heavy bottleneck.
- (xviii) Under strong driver speed adaptation, a driver that approaches a slower-moving preceding vehicle adapts the speed to that of the preceding vehicle quickly. In this case, all wide moving jams can dissolve, and only synchronized flow remains in congested traffic. The effect of moving jam dissolution can be called the “jam-absorption” effect or the effect of moving jam dissolution in synchronized flow. If through the use of ITS conditions of strong driver speed adaptation are realized in congested traffic, all moving jams dissolve in congested traffic. This can considerably decrease fuel consumption in a traffic network together with the increase in traffic safety.

Future Directions

Although many empirical macroscopic spatiotemporal congested pattern features have already been understood, there are still many problems in understanding empirical microscopic features of these patterns as well as phase transitions within congested patterns, in particular occurring on highway sections with many adjacent bottlenecks as well as in city traffic.

Many of the microscopic empirical features of traffic breakdown at traffic signals in city traffic that have been predicted recently have not been studied sufficiently.

Possible induced phase transitions and possible complex pattern transformations caused by upstream propagation of congested patterns onto neighborhood sections of a traffic network have not been studied sufficiently.

Traffic flow models in the framework of the three-phase theory, which can explain all known empirical and experimental spatiotemporal features of traffic breakdown and congested traffic, are only at the beginning of their development. There are almost no analytical studies of the spatiotemporal characteristics of congested patterns within the three-phase theory.

An empirical microscopic spatiotemporal dynamics of traffic congestion at heavy bottlenecks has not still been studied sufficiently.

Vehicle systems for automated (autonomous) driving as well as other ITS applications in the framework of the three-phase theory that can help to increase safety and comfort while driving in congested traffic are only at the beginning of their development.

Other important questions, which should be answered in more details with the use of the three-phase theory are as follows:

1. What are empirical features of congested patterns that can be influenced for reliable spatial limitation of congestion growth and/or for congestion dissolution?
2. How can driver behavior change features of congested patterns with the aim of the increase safety and comfortable driving in congested traffic?
3. What are vehicle systems for automated driving and other ITS applications that can help to increase safety and comfort while driving in congested traffic?

These and many other unsolved problems are interesting and important fields of the future empirical and theoretical investigations of spatiotemporal congested traffic patterns.

Acknowledgements I would like to thank Sergey Klenov for the help and useful suggestions. We thank our partners for their support in the project "MEC-View – Object detection for automated driving based on Mobile Edge

Computing," funded by the German Federal Ministry of Economic Affairs and Energy.

Bibliography

- Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modeling. *Math Mod Methods App Sci* 12:1801–1843
- Chandler RE, Herman R, Montroll EW (1958) Traffic dynamics: studies in car following. *Oper Res* 6:165–184
- Chen D, Laval JA, Ahn S, Zheng Z (2012a) Microscopic traffic hysteresis in traffic oscillations: a behavioral perspective. *Transp Res B* 46:1440–1453
- Chen D, Laval JA, Zheng Z, Ahn S (2012b) A behavioral car-following model that captures traffic oscillations. *Transp Res B* 46:744–761
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
- Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
- Daganzo CF (1997) *Fundamentals of transportation and traffic operations*. Elsevier Science Inc., New York
- Davis LC (2004) Multilane simulations of traffic phases. *Phys Rev E* 69:016108
- Davis LC (2006a) Controlling traffic flow near the transition to the synchronous flow phase. *Physica A* 368:541–550
- Davis LC (2006b) Effect of cooperative merging on the synchronous flow phase of traffic. *Physica A* 361:606–618
- Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. *Physica A* 379:274–290
- Dion F, Rakha H, Kang YS (2004) Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transp Res B* 38:99–122
- Edie LC (1961) Car-following and steady state theory for non-congested traffic. *Oper Res* 9:66–77
- Edie LC, Foote RS (1958) Traffic flow in tunnels. *Highway Res Board Proc Ann Meeting* 37:334–344
- Edie LC, Foote RS (1960) Effect of shock waves on tunnel traffic flow. In: *Highway research board proceedings*, 39. HRB, National Research Council, Washington, DC, pp 492–505
- Edie LC, Herman R, Lam TN (1980) Observed multilane speed distribution and the kinetic theory of vehicular traffic. *Transp Sci* 14:55–76
- Elefteriadou L (2014) *An introduction to traffic flow theory*. Springer optimization and its applications, vol 84. Springer, Berlin
- Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Transp Res Rec* 1484:80–89

- Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) *Traffic and granular flow'01*. Springer, Heidelberg
- Gao K, Jiang R, Hu S-X, Wang B-H, Wu Q-S (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. *Phys Rev E* 76:026105
- Gartner NH, Stamatiadis C (2009) Traffic networks, optimization and control of urban. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9470–9500
- Gartner NH, Messer CJ, Rath A (eds) (1997) *Special Report 165: Revised Monograph on Traffic Flow Theory*. Transportation Research Board, Washington, DC
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7:499–505
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
- Geroliminis N, Skabardonis A (2011) Identification and analysis of queue spillovers in city street networks. *IEEE Trans Intell Transp Syst* 12:1107–1115. <https://doi.org/10.1109/TITS.2011.2141991>
- Haight FA (1963) *Mathematical theories of traffic flow*. Academic Press, New York
- Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Trans Res Rec* 1320:91–98
- Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transp Res Rec* 1365:12–18
- Hausken K, Rehborn H (2015) Game-theoretic context and interpretation of Kerner's three-phase traffic theory. In: Hausken K, Zhuang J (eds) *Game theoretic analysis of congestion, safety and security*. Springer series in reliability engineering. Springer, Berlin, pp 113–141
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) (2000) *Traffic and granular flow'99*. Springer, Heidelberg
- Hemmerle P, Koller M, Hermanns G, Rehborn H, Kerner BS, Schreckenberg M (2016a) Impact of synchronized flow in oversaturated city traffic on energy efficiency of conventional and electrical vehicles. *Collective Dyn* 1, A7:1–27. <https://doi.org/10.17815/CD.2016.7>
- Hemmerle P, Koller M, Rehborn H, Kerner BS, Schreckenberg M (2016b) Fuel consumption in empirical synchronized flow in urban traffic. *IET Intell Transp Syst* 10:122–129. <https://doi.org/10.1049/iet-its.2015.0014>
- Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7:86–106
- Hermanns G, Hemmerle P, Rehborn H, Koller M, Kerner BS, Schreckenberg M (2015) Microscopic simulation of synchronized flow in oversaturated city traffic: effect of drivers speed adaptation. *Transp Res Rec* 2490:47–55
- Hoogendoorn SP, Luding S, Bovy PHL, Schreckenberg M, Wolf DE (eds) (2005) *Traffic and granular flow'03*. Springer, Heidelberg
- Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. *J Phys A Math Gen* 37:8197–8213
- Jiang R, Wu QS (2005) Toward an improvement over Kerner-Klenov-Wolf three-phase cellular automaton model. *Phys Rev E* 72:067103
- Jiang R, Wu QS (2007) Dangerous situations in a synchronized flow model. *Physica A* 377:633–640
- Jiang R, Hu M-B, Wang R, Wu Q-S (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner-Klenov speed adaptation model. *Phys Lett A* 365:6–9
- Jiang R, Hu M-B, Zhang HM, Gao Z-Y, Jia B, Wu Q-S, Wang B, Yang M (2014) Traffic experiment reveals the nature of car-following. *PLOS One* 9:e94351
- Jiang R, Hu M-B, Zhang HM, Gao ZY, Jia B, Wu QS (2015) On some experimental features of car-following behavior and how to model them. *Transp Res B* 80:338–354
- Jiang R, Jin C-J, Zhang HM, Huang Y-X, Tian J-F, Wang W, Hu M-B, Jia B (2017) Experimental and empirical investigations of traffic flow instability. *Transp Res Proc* 23:157–173
- Jiang R, Jin C-J, Zhang HM, Huang Y-X, Tian J-F, Wang W, Hu M-B, Wang H, Jia B (2018, in press) Experimental and empirical investigations of traffic flow instability. *Transp Res C*. <https://doi.org/10.1016/j.trc.2017.08.024>
- Kaufmann S, Kerner BS, Rehborn H, Koller M, Klenov SL (2017) Aerial observation of inner city traffic and analysis of microscopic data at traffic signals. *TRB Paper* 17-03078. Transportation research board 96th annual meeting, Washington, DC
- Kaufmann S, Kerner BS, Rehborn H, Koller M, Klenov SL (2018) Aerial observations of moving synchronized flow patterns in over-saturated city traffic. *Transp Res C* 86:393–406
- Kerner BS (1998a) Theory of congested traffic flow. In: Rysgaard R (ed) *Proceedings of the 3rd symposium on highway capacity and level of service*, vol 2. Road Directorate, Ministry of Transport, Denmark, pp 621–642
- Kerner BS (1998b) Traffic flow: experiment and theory. In: Schreckenberg and Wolf (1998), pp 239–267
- Kerner BS (1998c) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
- Kerner BS (1999a) Congested traffic flow: observations and theory. *Trans Res Rec* 1678:160–167
- Kerner BS (1999b) Theory of congested traffic flow: self-organization without bottlenecks. In: Ceder A (ed) *Transportation and traffic theory*. Elsevier Science, Amsterdam, pp 147–171
- Kerner BS (1999c) The physics of traffic. *Phys World* 12:25–30

- Kerner BS (2000a) Phase transitions in traffic flow. In: Helbing et al. (2000), pp 253–284
- Kerner BS (2000b) Experimental features of the emergence of moving jams in free traffic flow. *J Phys A Math Gen* 33:L221–L228
- Kerner BS (2001) Complexity of synchronized flow and related problems for basic assumptions of traffic flow theories. *Netw Spat Econ* 1:35–76
- Kerner BS (2002a) Synchronized flow as a new traffic phase and related problems for traffic flow modelling. *Math Comput Model* 35:481–508
- Kerner BS (2002b) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
- Kerner BS (2004a) Three-phase traffic theory and highway capacity. *Physica A* 333:379–440
- Kerner BS (2004b) *The physics of traffic*. Springer, Berlin/New York
- Kerner BS (2007) Features of traffic congestion caused by bad weather conditions and accidents. E-print arXiv:0712.1728; Available at <http://arxiv.org/abs/0712.1728>
- Kerner BS (2008) A theory of traffic congestion at heavy bottlenecks. *J Phys A Math Theor* 41:215101
- Kerner BS (2009a) Traffic congestion, modelling approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9302–9355
- Kerner BS (2009b) Traffic congestion, spatiotemporal features of. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9355–9411
- Kerner BS (2009c) *Introduction to modern traffic flow theory and control*. Springer, Berlin/New York
- Kerner BS (2011) Physics of traffic gridlock in a city. *Phys Rev E* 84:045102(R)
- Kerner BS (2012) Complexity of spatiotemporal traffic phenomena in flow of identical drivers: explanation based on fundamental hypothesis of three-phase theory. *Phys Rev E* 85:036110
- Kerner BS (2013a) The physics of green-wave breakdown in a city. *Europhys Lett* 102:28010
- Kerner BS (2013b) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Physica A* 392:5261–5282
- Kerner BS (2014) Three-phase theory of city traffic: moving synchronized flow patterns in under-saturated city traffic at signals. *Physica A* 397:76–110
- Kerner BS (2015a) Microscopic theory of traffic-flow instability governing traffic breakdown at highway bottlenecks: growing wave of increase in speed in synchronized flow. *Phys Rev E* 92:062827
- Kerner BS (2015b) Failure of classical traffic flow theories: a critical review. *Elektrotechnik und Informationstechnik* 132:417–433
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Physica A* 450:700–747
- Kerner BS (2017a) *Breakdown in traffic networks: fundamentals of transportation science*. Springer, Berlin/New York
- Kerner BS (2017b) Traffic networks, breakdown in. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. Springer, Berlin. https://doi.org/10.1007/978-3-642-27737-5_701-1
- Kerner BS (2018a) Traffic breakdown, modeling approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. Springer, Berlin. https://doi.org/10.1007/978-3-642-27737-5_559-2
- Kerner BS (2018b) Physics of automated driving in framework of three-phase traffic theory. *Phys Rev E* 97:042303
- Kerner BS (2018c) Autonomous driving in the framework of three-phase traffic theory. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. Springer, Berlin
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35:L31–L43
- Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68:036130
- Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A Math Gen* 39:1775–1809
- Kerner BS, Klenov SL (2009) Phase transitions in traffic flow on multilane roads. *Phys Rev E* 80:056101
- Kerner BS, Klenov SL (2010) A theory of traffic congestion at moving bottlenecks. *J Phys A Math Theor* 43:425101
- Kerner BS, Konhäuser P (1994) Structure and parameters of clusters in traffic flow. *Phys Rev E* 50:54–83
- Kerner BS, Rehborn H (1996a) Experimental features and characteristics of traffic jams. *Phys Rev E* 53:R1297–R1300
- Kerner BS, Rehborn H (1996b) Experimental properties of complexity in traffic flow. *Phys Rev E* 53:R4275–R4278
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. *Phys Rev Lett* 79:4030–4033
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35:9971–10013
- Kerner BS, Klenov SL, Hiller A (2006a) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. *J Phys A Math Gen* 39:2001–2020
- Kerner BS, Klenov SL, Hiller A, Rehborn H (2006b) Microscopic features of moving traffic jams. *Phys Rev E* 73:046107
- Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. *Non Dyn* 49:525–553
- Kerner BS, Klenov SL, Hermanns G, Hemmerle P, Rehborn H, Schreckenberg M (2013a) Synchronized flow in oversaturated city traffic. *Phys Rev E* 88:054801
- Kerner BS, Rehborn H, Schäfer R-P, Klenov SL, Palmer J, Lorkowski S, Witte N (2013b) Traffic dynamics in empirical probe vehicle data studied with three-phase theory: spatiotemporal reconstruction of traffic phases and generation of jam warning messages. *Physica A* 392:221–251

- Kerner BS, Hemmerle P, Koller M, Hermanns G, Klenov SL, Rehborn H, Schreckenberg M (2014a) Empirical synchronized flow in oversaturated city traffic. *Phys Rev E* 90:032810
- Kerner BS, Klenov SL, Schreckenberg M (2014b) Traffic breakdown at a signal: classical theory versus the three-phase theory of city traffic. *J Stat Mech* P03001
- Kerner BS, Koller M, Klenov SL, Rehborn H, Leibel M (2015) The physics of empirical nuclei for spontaneous traffic breakdown in free flow at highway bottlenecks. *Physica A* 438:365–397
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: microscopic description of traffic phases. *Phys Rev E* 65:056133
- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. *Phys Rev E* 70:016115
- Koshi M, Iwasaki M, Ohkura I (1983) Some findings and an overview on vehicular flow characteristics. In: Hurdle VF (ed) *Proceedings of 8th international symposium on transportation and traffic theory*. University of Toronto Press, Toronto, p 403
- Laval JA (2007) Linking synchronized flow and kinematic waves. In: Schadschneider et al. (2007), pp 521–526
- Laval JA, Toth CS, Zhou Y (2014) A parsimonious model for the formation of oscillations in car-following models. *Transp Res B* 70:228–238
- Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical restriction versus human overreaction triggering congested traffic states. *Phys Rev Lett* 92:238702
- Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
- Mahmassani HS (ed) (2005) *Transportation and traffic theory*. In: *Proceedings of the 16th international symposium on transportation and traffic theory*. Elsevier, Amsterdam
- Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
- May AD (1990) *Traffic flow fundamentals*. Prentice-Hall, Inc., New Jersey
- Molzahn S-E, Kerner BS, Rehborn H, Klenov SL, Koller M (2017) Analysis of speed disturbances in empirical single vehicle probe data before traffic breakdown. *IET Intell Transp Syst* 11:604–612. <https://doi.org/10.1049/iet-its.2016.0315>
- Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
- Nagel K, Wagner P, Woessler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–716
- Nakayama A, Fukui M, Kikuchi M, Hasebe K, Nishinari K, Sugiyama Y, Tadaki S-I, Yukawa S (2009) Metastability in the formation of an experimental traffic jam. *New J Phys* 11:083025
- Neubert L, Santen L, Schadschneider A, Schreckenberg M (1999) Single-vehicle data of highway traffic: a statistical analysis. *Phys Rev E* 60:6480–6490
- Newell GF (1982) *Applications of queuing theory*. Chapman Hall, London
- Next Generation Simulation Programs. <http://ops.fhwa.dot.gov/trafficanalysis/tools/ngsim.htm>. Accessed 25 Sept 2016
- Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Trans Res Rec* 1634:64–69
- Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical restriction versus human overreaction: accident avoidance and two-lane simulations. In: Schadschneider et al. (2007), pp 503–508
- Rehborn H, Klenov SL (2009) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer, Berlin, pp 9500–9536
- Rehborn H, Koller M (2014) A study of the influence of severe environmental conditions on common traffic congestion features. *J Adv Transp* 48:1107–1120
- Rehborn H, Palmer J (2008) ASDA/FOTO based on Kerner's three-phase traffic theory in North Rhine-Westphalia and its integration into vehicles. In: *Intelligent vehicles symposium*. IEEE, pp 186–191
- Rehborn H, Klenov SL, Palmer J (2011a) An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. *Physica A* 390:4466–4485
- Rehborn H, Klenov SL, Palmer J (2011b) Common traffic congestion features studied in USA, UK, and Germany based on Kerner's three-phase traffic theory. In: *Intelligent vehicles symposium (IV)*. IEEE, pp 19–24
- Rehborn H, Klenov SL, Koller M (2017) Traffic prediction of congested patterns. In: Meyers RA (ed) *Encyclopedia of complexity and system science*. Springer Science+Business Media LLC. https://doi.org/10.1007/978-3-642-27737-5_564-2
- Rempe F, Franeck P, Fastenrath U, Bogenberger K (2016) Online freeway traffic estimation with real floating car data. In: *Proceedings of 2016 I.E. 19th international conference on ITS*, pp 1838–1843
- Rempe F, Franeck P, Fastenrath U, Bogenberger K (2017) A phase-based smoothing method for accurate traffic speed estimation with floating car data. *Trans Res C* 85:644–663
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and granular flow'05*. In: *Proceedings of the international workshop on traffic and granular flow*. Springer, Berlin
- Schreckenberg M, Wolf DE (eds) (1998) *Traffic and granular flow'97*. In: *Proceedings of the international workshop on traffic and granular flow*. Springer, Singapore
- Sugiyama Y, Fukui M, Kikuchi M, Hasebe K, Nakayama A, Nishinari K, Tadaki S, Yukawa S (2008) Traffic jam without bottleneck – experimental evidence for the physical mechanism of forming a jam. *New J Phys* 10:033001
- Tadaki S, Kikuchi M, Fukui M, Nakayama A, Nishinari K, Shibata A, Sugiyama Y, Yoshida T, Yukawa S (2013) Phase transition in traffic jam experiment on a circuit. *New J Phys* 15:103034
- Tadaki S, Kikuchi M, Nakayama A, Shibata A, Sugiyama Y, Yukawa S (2016) Characterizing and distinguishing free and jammed traffic flows from the distribution and correlation of experimental speed data. *New J Phys* 18:083022

- Tian J-F, Jiang R, Jia B, Gao Z-Y, Ma SF (2016a) Empirical analysis and simulation of the concave growth pattern of traffic oscillations. *Transp Res B* 93:338–354
- Tian J-F, Jiang R, Li G, Treiber M, Jia B, Zhu CQ (2016b) Improved 2D intelligent driver model in the framework of three-phase traffic theory simulating synchronized flow and concave growth pattern of traffic oscillations. *Transp Rec F* 41:55–65
- Tian J-F, Jiang R, Treiber M (2016c) Noise induced pattern formation of oscillation growth in traffic flow. arXiv:1611.00527. <http://arxiv.org/abs/1611.00527>
- Treiber M, Kesting A (2013) *Traffic flow dynamics*. Springer, Berlin
- Treiterer J (1975) Investigation of traffic dynamics by aerial photogrammetry techniques. Ohio State University Technical Report PB 246 094, Columbus
- Treiterer J, Myers JA (1974) The hysteresis phenomenon in traffic flow. In: Buckley DJ (ed) *Proceedings of 6th international symposium on transportation and traffic theory*. A.H. & AW Reed, London, pp 13–38
- Treiterer J, Taylor JI (1966) Traffic flow investigations by photogrammetric techniques. *Highway Res Rec* 142:1–12
- Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. *Physica A* 378:475–484
- Webster FV (1958) *Traffic signal settings*. HMSO, London
- Whitham GB (1974) *Linear and nonlinear waves*. Wiley, New York
- Wiedemann R (1974) *Simulation des Verkehrsflusses*. University of Karlsruhe, Karlsruhe
- Wolf DE (1999) Cellular automata for traffic simulations. *Physica A* 263:438–451
- Zeng J-W, Yang X-G, Qian Y-S, Wei X-T (2017) Research on three-phase traffic flow modeling based on interaction range. *Modern Phys Lett B* 1750328. <https://doi.org/10.1142/S0217984917503286>

Traffic Prediction of Congested Patterns

H. Rehborn¹, Sergey L. Klenov² and M. Koller¹

¹Research and Development RD/USN, HPC: 059-X901, Daimler AG, Sindelfingen, Germany

²Department of Physics, Moscow Institute of Physics and Technology, Moscow, Russia

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Congested Traffic Pattern Features on Freeways Relevant for Prediction

Detection of Congested Traffic Patterns Based on Probe Vehicles

Reconstruction of Synchronized Flow Patterns at Traffic Signal in City Traffic with the Use of Probe Vehicle Data

Fuel Consumption in Road Networks

Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data

Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner's Three-Phase Traffic Theory Versus Other Approaches

Long-Term Traffic Prediction

Applications of Traffic Prediction

Future Directions

Bibliography

Glossary

Data mining or physics of traffic: two different approaches to traffic prediction In “data mining” approaches, reproducible features of measured traffic data are identified and analyzed through the use of various learning systems like neural networks, regression techniques, time series, wavelet analysis, Bayesian networks, etc. The nature of the traffic phenomena does not play a key role for this reproducible traffic

feature learning. Instead, reproducible phenomena of the input data are learned by those methods. In contrast, “physics of traffic” approaches understand, explain, and model these reproducible features of measured traffic data. Then features of traffic phenomena are used as the basis for traffic prediction methods.

Fuel consumption Traffic congestions have a specific impact on the fuel and/or energy consumption of vehicles: this consumption is traffic state dependent on freeways and on urban roads. Predictions of the traffic states offer the chance to develop energy-efficient driving for vehicles and dynamic navigation on energy-efficient routes. In times of reduced resources, the precise traffic prediction and its energetic efforts in vehicles might become more and more important.

Kerner's three-phase traffic theory In Kerner's three-phase traffic theory, empirical (measured) features of spatiotemporal congested traffic patterns are explained. Three traffic phases are identified, and the features of spatiotemporal dynamics within these traffic phases, as well as the phase transitions between them, are examined. These three traffic phases are (i) free traffic, (ii) synchronized traffic, and (iii) wide moving jam. The synchronized flow and wide moving jam traffic phases are associated with a congested traffic state. These two traffic phases are defined through the following empirical criteria. The definition of the wide moving jam phase (J): a wide moving jam is a moving jam that propagates through any bottleneck while maintaining the mean velocity of the downstream jam front. The definition of the synchronized flow phase (S): the downstream front of a synchronized flow region does not exhibit the above characteristic jam feature; in particular, the downstream front of a synchronized flow is often fixed at the location of the bottleneck. The main reason for Kerner's three-phase traffic theory is the explanation of the empirical nucleation nature of traffic

breakdown at a highway bottleneck: in accordance with real field traffic data (empirical data), Kerner's three-phase theory explains traffic breakdown through a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) in a metastable free flow at the bottleneck.

Long-term prediction with time series Long-term prediction is defined to be independent from the current situation, i.e., traffic variables solely from a historic database are used. This is the established method for traffic prediction with larger time horizons, e.g., planning tomorrow's roadwork. In contrast, short-term traffic prediction is influenced by the current traffic situation. Without this influence, the most probable time series from the database can be chosen directly to perform a specific traffic variable prediction.

Measurements of traffic flow variables Traffic variables like flow rate (normally given as number of vehicles per hour), vehicle speed, and occupancy (percentage of time in which a road location (i.e., normally a detector) is occupied by a vehicle in a certain time interval) are measured with local detectors (single or double induction loop detectors) in the road network. The locally measured traffic variables are averaged and transferred from the roadside detectors to online traffic management centers in usually fixed time intervals (commonly 30 s in the USA, 60 s in Germany). Individual vehicles can measure travel times on road segments, which can be aggregated to travel times.

Onset of congestion Onset of congestion (traffic breakdown) is defined as congested traffic pattern emergence in an initial free flow. The onset of congestion is observed mostly at freeway bottlenecks. In Kerner's three-phase traffic theory, the onset of congestion is explained by a phase transition from free flow to synchronized traffic flow at a bottleneck ($F \rightarrow S$ transition for short).

Prediction of spatiotemporal congested traffic pattern The function of the prediction approach is to forecast the whole congestion picture in space and time. Parameters of the congested patterns, such as size and form, at specific bottlenecks can provide a "fingerprint" of the road network, because they are typical, recurring, and specific to the related bottleneck. These

features facilitate the predictability of traffic congestion.

Probe vehicle data After the removal of the selective availability in 2000, the GPS (Global Positioning System) offers a position accuracy in the order of 5–10 m: vehicles and all mobile devices can be localized on roads with this high accuracy. Connected vehicles and several mobile devices send their driving positions in small time intervals to central servers: the reconstruction of the traffic state is now based on masses of such probe vehicle data. The probe vehicle data give the opportunity to retrieve the current traffic situation on all roads and open, therefore, the window for traffic prediction in networks based on measurements.

Recurring and nonrecurring congestion The congestion emerging by the routine presence of large numbers of vehicles on roads is called recurring congestion. Unexpected disruptions of traffic caused by like lane blockings, incidents from work zones, weather events, etc. are called nonrecurring congestion. Spatiotemporal congested patterns are caused by both recurring and nonrecurring congestion. After recognition of a congested pattern, the further prediction is mainly independent from its original reason. The methods described in this contribution are valid for both kinds of congestion reasons.

Short-term prediction with time series A short-term prediction with time series can be performed via the comparison of a historical database of traffic variables and the current traffic data measurements. The choice of the appropriate traffic data curve that is then used for traffic prediction is called "matching." This choice is performed from the historical database finding the closest match to the current traffic data. The best matched time series is not necessarily the one with the highest probability among the different time series in the database. Therefore, matching can increase the prediction quality considerably.

Spatiotemporal congested traffic pattern Spatiotemporal congested traffic patterns are traffic patterns on freeways that emerge and develop in space and time as a result of the

onset of congestion in an initially free traffic flow. Spatiotemporal congested traffic patterns occur mostly at freeway bottlenecks associated with on- and off-ramps, roadwork, decreasing freeway lane numbers, etc.

Time series of traffic variables Traffic variables can be archived and combined as daily or hourly time series. For example, a flow rate time series could be used as the traffic demand estimation for vehicles driving into congestion. A time series of travel time for a road network could be used for dynamic route guidance applications. Important parameters for the choice of the most probable time series used for prediction are at least the characteristics of the specific day (e.g., Monday morning rush hour) and the probability of occurrence of each of the time series of traffic variables for days with the same characteristics.

Wide moving jam emergence In empirical observations, wide moving jams emerge spontaneously in synchronized flow. As explained by Kerner's three-phase traffic theory, this wide moving jam emergence is associated with a first-order phase transition from synchronized flow to wide moving jam ($S \rightarrow J$ transition for short). Thus, wide moving jams emerge in free flow as a result of a sequence of two-phase transitions: firstly, a phase transition at a certain freeway location from free flow to synchronized traffic flow occurs. Later and usually at another road location, this synchronized flow can transform into a spatiotemporal congested pattern with propagating wide moving jams.

Definition of the Subject and Its Importance

Empirical congested traffic patterns and their predictions are an interesting field of complex systems because of their complex spatiotemporal behavior. The physics of traffic with its nonlinear phenomena, self-organization processes based on individual driver behavior, and phase transitions from free to congested traffic states are a great challenge for traffic prediction approaches. The

road infrastructure approaches its capacity with severe congestion problems because of the high and increasing mobility demand in many countries of the world: a better understanding of traffic congestion patterns, their emergence, and their dissolution as well as their prediction may improve both the complex individual and collective management of traffic networks. Due to advancing traffic data measurement techniques, the variety and complexity of available traffic data are steadily increasing. The amount of available traffic data can be used as archive information for the time series of traffic variables. Traffic predictions in the long term need historical knowledge about the traffic variables. The available historical traffic data varies from single-vehicle trajectory data and large amounts of local induction loop detector data up to traffic message databases from service broadcasters. Specifically, the traffic flow rate and/or the travel times are needed for an estimation of future traffic situation. In the long-term, traffic patterns are recurring and regular (like a morning peak traffic hour with commuters). This regularity of traffic makes long-term time series-based traffic predictions possible and feasible. On the other hand, a comparison with a current situation is a basis of short-term traffic prediction.

Relevant applications of traffic predictions are, from the collective point of view, the network predictions of traffic states as basic information for traffic management and control as well as the planning of the probable congestion consequences of roadwork. From an individual driver's point of view, traffic prediction is important for efficient route choice and the estimation of arrival times. Additionally, precise congestion information could be used for safety and comfort applications in vehicles like the adaptation of cruise control systems.

In the near future, automated or semiautomated vehicles will drive in the road networks: essentially, these vehicles without driver control need predictive information of all events in their surroundings, e.g., congestions, dangers on the road, environmental issues, etc. For all automated driving the vehicle would need a predicted trajectory for its driving which must be based on the

predicted traffic on the road ahead. As one of the closer next development steps, automated driving in congested patterns has a high probability of realization due to the lower speeds and relatively deterministic controllable environment in a congestion: one comfort function of a vehicle could be the automated driving during congestion, while the driver does what he wants rather than driving by himself.

Introduction

Every morning, the commuters in urban areas of the world suffer from time delays caused by traffic congestion. The congestion is regular, recurring, and permanent at certain locations of the freeway road network. Therefore, from a better understanding of the underlying physical processes in traffic, the wish for precise and reliable traffic prediction arises for multiple traffic management purposes, both collective (e.g., traffic control) or individual (e.g., dynamic route guidance). To reach these goals, empirical congestion patterns must first be analyzed and understood correctly.

Throughout the modern world, traffic congestion has enormous associated costs in terms of delay, fuel consumption, emissions, etc. One way to alleviate this problem is to build new transportation facilities and expand the transportation networks which is very cost intensive and not always possible. Another cost-effective method would be the implementation of real-time and predicted traffic generated and disseminated to the traveling public for both pre-trip planning and en route usage. Having accurate and up-to-date traffic information, travelers would be able to have the opportunity to plan their trip and adjust their routes accordingly.

Infrastructure planning needs information on the traffic demand for its long-term construction processes. Additionally, the planning of roadwork must ensure efficiency, minimizing the effect on the current traffic demand while also considering requirements regarding costs and working necessities. The individual driver is interested in long-term traffic information for his pre-trip planning which can range from tomorrow's trip to work to

his next holiday journey further in the future. Long-term time series of traffic based on historical data are an opportunity to fulfill such individual or collective wishes for forecasted traffic information. The goal of the prediction approach is the precise prediction of a traffic variable like flow rate or travel time. Next, the probability of traffic breakdown, i.e., the possible onset of congestion at a certain location in the road network, should be predicted. After this congestion emergence, the spatiotemporal behavior of the congested region must be forecasted.

Figure 1 illustrates a possible qualitative scheme to perform a traffic prediction: historical time series of traffic variables and the reconstruction of the current traffic state are the "input data" for traffic prediction methods including "matching," i.e., a choice of the most probable traffic pattern, which are used then for the traffic prediction of traffic variables.

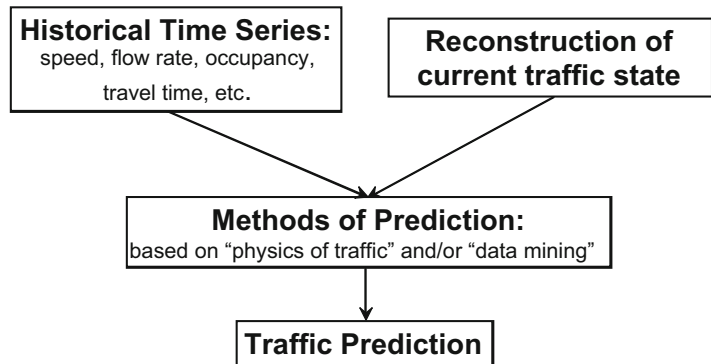
There are a huge number of various approaches to reconstruction of current traffic state and short-term prediction (Fig. 1). Related approaches can be subdivided into two main categories, distinguished by their principal concept: driven by "physics of traffic" or "data mining" (Fig. 2). Some approaches, as presented in this article, combine both "physics of traffic" and "data mining."

In the article, the data mining approaches are illustrated with regard to their prediction capability; in particular, we discuss the following methods for traffic prediction: "statistical regression," "neural networks," "wavelet techniques," and "cluster algorithms." It should be noted that data mining technologies may be used in addition for the generation of databases of historical time series.

The approaches associated with "physics of traffic" are devoted to the understanding of empirical (measured) spatiotemporal features of traffic as well as to modeling of these features. Then these features are used for traffic prediction. Although there were a huge number of studies of measured data on freeways (see classic works, e.g., by Edie and Foote (1960), Treiterer (1975), and Koshi et al. (1983), and references in the book (Kerner 2004)), only recently Kerner solved the puzzle of empirical spatiotemporal features of freeway traffic congestion (e.g., Kerner 1998,

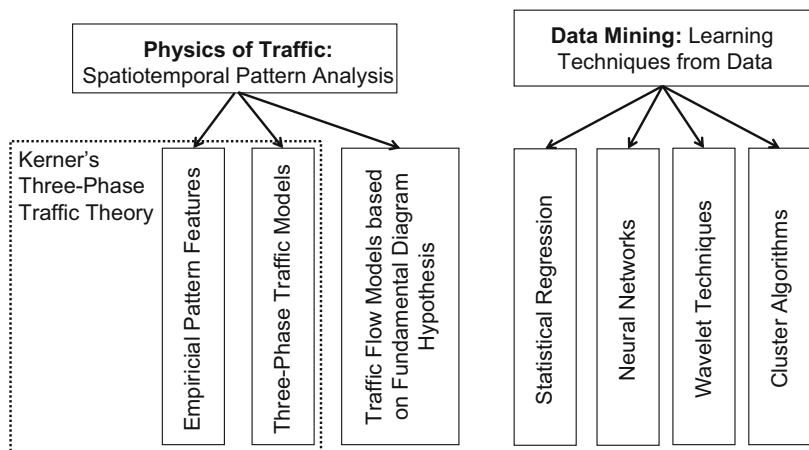
Traffic Prediction of Congested Patterns,

Fig. 1 Possible qualitative scheme of traffic prediction



Traffic Prediction of Congested Patterns,

Fig. 2 General overview on approaches to reconstruction of current traffic state and short-term prediction



1999c, 2004, 2008, 2009; Kerner and Herrtwich 2001; Kerner and Rehborn 1996a, b, 1997). Kerner's three-phase traffic theory has been developed to explain the puzzle of freeway measured data.

It turns out that earlier known modeling approaches to freeway traffic congestion (e.g., Daganzo 1994, 1995, 1999; Helbing 1997; May 1990; Nagel and Schreckenberg 1992; Schönhof and Helbing 2007; "Traffic Flow Theory 2006" 2006), which are based on the fundamental-diagram hypothesis (Fig. 2), cannot explain and predict the empirical probabilistic features of the onset of congestion and resulting congested patterns occurring on freeways that are required for traffic prediction (see Kerner 2007, 2008, 2009; Kerner and Herrtwich 2001 for more detail). These traffic flow models are standard ones for validation of freeway traffic prediction. Thus, the

related simulations or other applications of these models cannot be used for a reliable freeway traffic prediction. In other words, at time Kerner's three-phase traffic theory is the only traffic theory, which can explain the observed spatiotemporal phenomena of congested freeway traffic. For this reason, we will use in this article often Kerner's three-phase traffic theory (Kerner 2004) as the basic approach associated with "physics of traffic" application for traffic prediction on freeways.

Empirical features of freeway traffic that make the prediction difficult and complex are the probabilistic nature of the traffic breakdown and the probabilistic features of the spatiotemporal patterns of freeway traffic (Kerner 1998, 1999c, 2004, 2008, 2009; Kerner and Herrtwich 2001; Kerner and Rehborn 1996a, b, 1997). As one result to be explained later, the travel times could be predicted based on measured and archived

travel times in principle precisely only in free traffic situations which are the less interesting from the driver perspective because his personal trip would not be influenced by congestion. Possible applications of freeway traffic predictions are the use in navigation systems, for safety and comfort functions in vehicles and on the infrastructure side for the traffic assignment in networks and the choice of appropriate traffic control strategies.

The article starts with a description of the variety of congested traffic pattern on freeways including the main types of congested patterns and a brief history of the three-phase traffic theory (section “[Congested Traffic Pattern Features on Freeways Relevant for Prediction](#)”). The next section introduces the use of connected “probe” vehicles sending permanently their position data to central servers: based on huge amounts of such probe data, a prediction of dangerous jam tails is possible. In addition, two types of congested patterns at traffic signals (e.g., moving queues and synchronized flow) are measured based on probe data (section “[Reconstruction of Synchronized Flow Patterns at Traffic Signal in City Traffic with the Use of Probe Vehicle Data](#)”). Based on traffic predictions, the fuel and/or energy consumption of vehicles can be estimated (section “[Fuel Consumption in Road Networks](#)”). FOTO and ASDA models for reconstruction, tracking, and prediction of spatiotemporal congested pattern dynamics are the subject of section “[Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data](#).” These methods are also used in section “[Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner’s Three-Phase Traffic Theory Versus Other Approaches](#)” in which various approaches to reconstruction of current traffic state and short-term prediction (Fig. 2) are discussed. A long-time traffic prediction based on historical time series derived from measured traffic data are considered in section “[Long-Term Traffic Prediction](#).” In section “[Applications of Traffic Prediction](#),” some currently known applications of traffic prediction are explained. Finally, we discuss some future directions and remaining tasks in the field followed by an extended bibliography.

Congested Traffic Pattern Features on Freeways Relevant for Prediction

Definition of Traffic Phases and Some Hypotheses of Kerner’s Three-Phase Traffic Theory

After extensive traffic data analyses of available stationary measurements spanning several years, Kerner discovered that in congested freeway traffic, two different traffic phases must be differentiated: “synchronized flow” and “wide moving jam” (Kerner 1998, 1999c, 2002, 2004; Kerner and Rehborn 1996a, b, 1997). Experimental features of traffic on freeways were investigated and interpreted intensively in the late 1990s (Kerner 1999c; Kerner and Rehborn 1996a, b, 1997 and references in Kerner 2004). Empirical macroscopic spatiotemporal objective criteria for the phases as important elements of Kerner’s three-phase traffic theory (see the books Kerner 2004, 2009 for comprehensive explanations) are as follows:

- (i) A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or freeway bottlenecks.
- (ii) In contrast, the downstream front of the “synchronized flow” phase is often fixed at a freeway bottleneck. Within the front, vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow.

It must be noted that the observation of speed synchronization in congested traffic is no criterion for the phase differentiation as well as some other relationships and features of congested traffic measured at a freeway location (e.g., in the flow-density plane). The clear differentiation between the synchronized flow and wide moving jam phases can be made on these above objective criteria (i) and (ii) *only*. These objective criteria define the features of the two congested traffic phases (Kerner 2004). Traffic phase definition in congested traffic cannot be made definitely and consistently via differences in vehicle speeds, densities, or vehicle flow

rates or flow-density criteria in measured traffic variables.

As a consequence, the identification of the location and of the parameters of both traffic phases on the freeway is the basis for successful traffic prediction: the objective criteria have to be used to identify the traffic phases. After this identification, both traffic phases have predictable features and interdependences among each other which will be used in the prediction method. Some elements and issues from Kerner's three-phase traffic theory will introduce those predictable features. The aspects described here focus on the relevant features for traffic prediction.

Figure 3a illustrates a vehicle speed profile over time and location with measured traffic data and shows the three traffic phases. A wide moving jam propagates as a "low-speed valley" through the freeway stretch. In contrast, a second speed valley is almost fixed at the bottleneck location in Fig. 3a: this congested traffic phase belongs to the synchronized flow phase.

One of the important hypotheses of Kerner's three-phase traffic theory (Kerner 2004) is as follows: steady states of synchronized flow cover a two-dimensional (2D) region in the flow-density plane (so-called Kerner's 2D states) (Fig. 3b). This means that in contrast with the hypotheses of all earlier traffic flow theories, there is no fundamental diagram for steady states of traffic flow in this theory. The red line J (so-called Kerner's line J, Maerivoet and De Moor 2005) represents the velocity of the downstream front of a wide moving jam (Fig. 3b).

Some other hypotheses and results of the theory, at any density of free flow at which both an $F \rightarrow S$ transition and an $F \rightarrow J$ transition are possible, a critical speed (density) disturbance needed for the $F \rightarrow S$ transition is considerably smaller than that needed for the $F \rightarrow J$ transition. This should explain why empirical traffic breakdown is associated with a $F \rightarrow S$ transition. This explains also why despite of free flow metastability with respect to a $F \rightarrow J$ transition empirically observed, this $F \rightarrow J$ transition does not occur spontaneously in empirical observations.

One of the effects of the hypothesis of the three-phase traffic theory is as follows. When a

vehicle approaches the preceding vehicle that moves with a slower time-independent speed and cannot pass it, the vehicle adapts its own speed to the speed of the preceding vehicle at a space gap, which is only one possible space gap from the infinite amount of space gaps within the above 2D region at this speed (Fig. 3b): there is no desired (or optimal) space gap at a time independent, i.e., constant speed in the car-following behavior in this theory. This feature is called the speed adaptation effect in synchronized flow.

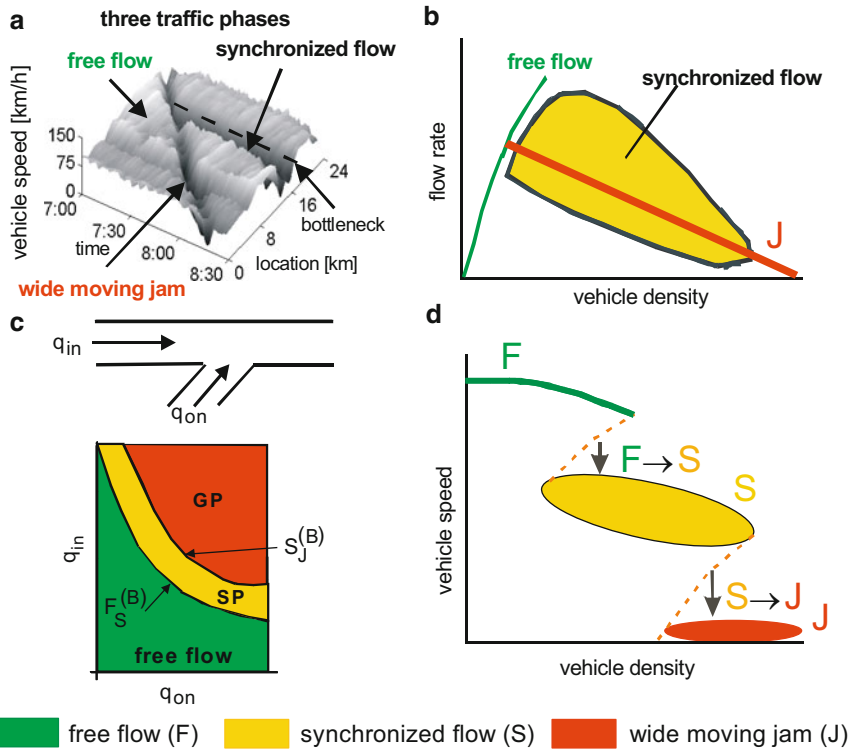
Traffic breakdown ($F \rightarrow S$ transition) is explained in Kerner's three-phase traffic theory as a competition between the speed adaptation effect and a well-known over-acceleration effect, i.e., a tendency of a vehicle to accelerate to a free flow speed when it seems to be possible.

All steady states on the line J (an infinite number) are threshold states for wide moving jam existence and emergence. The line J separates all steady states of synchronized flow in the flow-density plane into two different classes: all states below the line J are stable with respect to wide moving jam existence and emergence. All states on and above the line J are metastable states with respect to wide moving jam existence and emergence (Fig. 3b).

The simplified phase diagram at an on-ramp bottleneck illustrates the phase transitions. While increasing q_{on} (the flow rate to the on-ramp) at a same level of q_{in} (the flow rate on the main road) in free flow (Fig. 3c), the following occurs: the onset of congestion in an initial free flow at a freeway bottleneck is associated with a phase transition from free flow to synchronized flow. Firstly, synchronized flow emerges ($F \rightarrow S$ transition) at the bottleneck. The boundary $F_S^{(B)}$ is associated with this $F \rightarrow S$ transition. Wide moving jams can emerge spontaneously only in synchronized flow ($S \rightarrow J$ transition). The boundary $S_J^{(B)}$ is associated with the $S \rightarrow J$ transition (Fig. 3c). Thus, wide moving jams emerge due to a sequence of two-phase transitions called $F \rightarrow S \rightarrow J$ transitions.

A double Z-characteristic in the speed-density plane illustrates the sequence of the $F \rightarrow S \rightarrow J$ transitions and explains these phase transitions (Kerner 2004). The first Z-characteristic related to higher speeds explains the $F \rightarrow S$ transition (arrow labeled $F \rightarrow S$ in Fig. 3d) and includes

Kerner's Three-Phase Traffic Theory



Traffic Prediction of Congested Patterns, Fig. 3 Explanations to Kerner's three-phase traffic theory: (a) phase definition in measured data, (b) two-dimensional region of hypothetical steady states of synchronized flow, (c) simplified traffic phase diagram at an on-ramp bottleneck, and

(d) qualitative representation of the double Z-characteristic for phase transitions between free flow, synchronized flow, and wide moving jams (with low-speed traffic states within a wide moving jam (Kerner 2004, 2009))

the traffic states F and S. The second Z-characteristic, which is related to lower speeds, explains the $S \rightarrow J$ transition (arrow labeled $S \rightarrow J$ in Fig. 3d) and includes the traffic states S and J. Dashed curves in the double Z-characteristic are associated with the so-called critical speeds within local disturbances in traffic flow required for a phase transition (see for more details the books Kerner 2004, 2009). It must be noted that the double Z-characteristic of the three-phase traffic theory (Fig. 3d) can usually be correctly analyzed only after spatiotemporal features of the phase transitions and congested patterns have been studied. This is because $F \rightarrow S$ and $S \rightarrow J$ transitions occur at different locations.

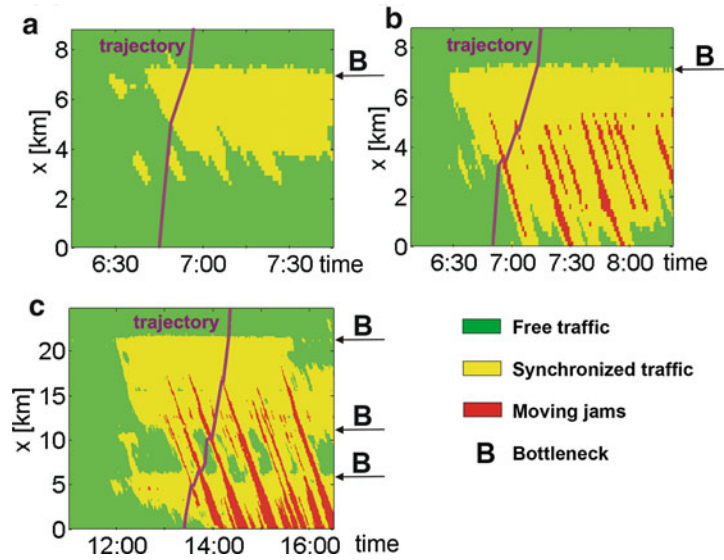
Main Types of Spatiotemporal Congested Traffic Patterns

Considering the case of an isolated effectual bottleneck (a bottleneck, which is far enough from other bottlenecks), the following two main congested pattern types can emerge (Kerner 2004):

1. "Synchronized pattern" (SP): this consists exclusively of an area of synchronized flow upstream of the effectual bottleneck, i.e., no wide moving jams develop in this pattern (Fig. 4a). In its variants, SP could be localized at the bottleneck ("LSP"), moving upstream ("MSP") or increasing over time ("WSP").
2. "General pattern" (GP): this consists of an area of synchronized flow upstream from the

Traffic Prediction of Congested Patterns,

Fig. 4 Spatiotemporal traffic pattern classification: (a) empirical SP (synchronized pattern), (b) empirical GP (general pattern), (c) empirical EP (expanded pattern) (Kerner 2004). Solid lines labeled “trajectory” illustrate some vehicle trajectories through these congested patterns



effectual bottleneck and wide moving jams, which develop in this synchronized flow spontaneously. GP covers thus both phases of congested traffic. It is the most frequently arising pattern at isolated bottlenecks on freeways (Fig. 4b).

If several effectual bottlenecks are closely neighboring, an expanded traffic pattern can develop (EP: expanded pattern) with areas of synchronized flow affecting several adjacent effectual bottlenecks (Fig. 4c). For each effectual bottleneck or for each set of several neighboring effectual bottlenecks, the spatiotemporal traffic patterns possess predictable, i.e., characteristic, and recurring characteristics, e.g., the most frequently arising traffic patterns or the average expansion of synchronized flow in a GP or EP. These characteristics can be almost alike on different days and years. The existence of traffic patterns also remain preserved within a large range of the flow rate. The prediction approach based on three-phase traffic theory uses this pattern definition and pattern features for the traffic pattern database concept. The pattern classification scheme is the key element of the related prediction approach. One of the main differences to other known approaches is the definition and

use of different traffic phases including their specific features and not based on aggregated information from freeways (e.g., “traffic states”) or drivers (e.g., “travel times”).

History of the Emergence of Kerner’s Three-Phase Traffic Theory

Historically, the first major impulse to the emergence of Kerner’s three-phase theory was made in 1995, when Kerner gave a talk “traffic moving jam without obvious reason” at the traffic engineering consulting company Heusch/Boesefeldt GmbH in Aachen (Germany) where more than 120 engineers work in traffic technology. In this talk, simulations of one of the classical traffic flow models (a version of Payne macroscopic traffic flow model) made by Kerner and Konhäuser (1994) have been considered. After that, in a discussion with the company founders, Jochen Boesefeldt and Heinz Heusch have stated that for more than 25 years the engineers have tried to implement ideas of classical traffic flow theory as well as traffic control methods based on these theories in real traffic installations: none of the methods and theories are in agreement with real field traffic data processed and stored by Heusch/Boesefeldt. As an example (Kirschfink et al. 2000; Kirschfink 1999) describe the use of complex traffic flow models with real traffic data: all the

classical models fail to reconstruct and predict traffic breakdown and resulting spatiotemporal traffic patterns in online installations. The failure of any applications of classical traffic flow theories for real traffic was the subject of a very intensive discussion of the traffic engineers with Boris Kerner during his stay at the Heusch/Boesefeldt GmbH (1995–1996). The work done together begins with an analysis of measured loop detector data provided by Ulrich Uerlings and Thomas Scheiderer on the freeway A5 in Hessen. Boris Kerner started with Hubert Rehborn a deep investigation of the available measured traffic data (Kerner and Rehborn 1996a, b, 1997) rather than to continue with simulations of classical traffic flow models.

The three-phase traffic theory introduced by Kerner between 1996 and 2002 answered the abovementioned question why the classical traffic flow theories are inconsistent with the reality: Kerner has found that real traffic breakdown at freeway bottleneck is a phase transition from free flow to synchronized flow ($F \rightarrow S$ transition) that exhibits a nucleation nature. Traffic breakdown occurs in a metastable state of free flow with respect to a $F \rightarrow S$ transition at the bottleneck: small enough disturbances in metastable free flow at a bottleneck decay. Contrarily, when a large enough disturbance in metastable free flow at the bottleneck occurs as a nucleus, traffic breakdown does occur at the bottleneck (Kerner 1999c, d). Empirical studies of a huge number of real field traffic data have fully confirmed this Kerner's conclusion: none of the classical traffic flow theories can show and explain the empirical nucleation nature of traffic breakdown at freeway bottlenecks (Kerner 2004, 2009, 2017). This fact explains why the classical traffic theories failed for application in the real world of traffic engineering.

Detection of Congested Traffic Patterns Based on Probe Vehicles

Since the year 2000, more and more vehicles and portable devices are connected and exchanging their data with central internet servers. The typical data are position data from the GPS (Global Positioning System) and the

related time stamp sent in cyclic intervals (together the GPS points build a trajectory of a probe vehicle). After removal of a Gaussian noise of the GPS signal – called “selective availability” – the raw GPS data has a position error von about 10 m only. Based on this accuracy, GPS raw data from vehicles and devices can be assigned to the related road segment (so-called matching). With map-matched probe vehicle data, we are able to derive information about the real-time traffic condition. Essentially, the map matching of trajectories to the underlying street network enables the ability to precisely reconstruct the traffic states on observed sections of a road. This gives the insights to improve traffic management, (highly) automated driving, and reliable control and optimization of traffic and transportation networks. A common way to visually show the traffic state are space-time diagrams (like in Fig. 4). In this section, we present the process of turning map-matched probe data (often called FCD (floating car data)) into space-time diagrams and give a brief interpretation of some examples.

One of the first detection and application of space-time diagrams can be seen in the research of traffic by Treiterer (1975). He observed traffic from aerial photography and reconstructed the traffic with tracking the vehicles with computer vision techniques. Figure 5 shows the time-space diagram on a short freeway section. Time-space diagrams show the trajectories (i.e., the position to any given time) of the cars on the observed road. They reveal the physical nature, structure, and characteristic parameters of a traffic jam. First patterns can be seen by the low-velocity wave propagating through the traffic.

After the matching process, the probe data of the vehicles are related spatial-temporally to a specific road stretch. Therefore, each single trajectory of a vehicle can be plotted over space and time. We follow the three-phase traffic theory (Kerner 2004, 2009) and draw the following general abstract diagram: over time, each vehicle has an individual speed profile (top of Fig. 6). If we draw the same vehicle trajectory over space and time (bottom of Fig. 6), we will see that the vehicle has passed two congested traffic phases

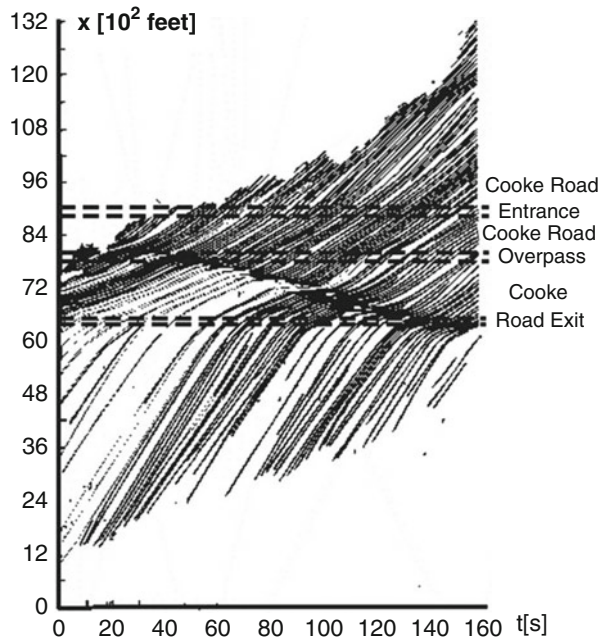
of Kerner’s theory: (i) wide moving jam and (ii) synchronized flow region.

For further analysis, we firstly introduce the term transition point for a phase transition between two different traffic phases along a trajectory of a probe vehicle and explain a qualitative difference between this transition point along a vehicle trajectory and

phase transition in traffic. A point for a phase transition between two different traffic phases along a trajectory of a probe vehicle gives a road location and a time instant at which the phase transition is just observed by a driver of the associated probe vehicle. This point is related to a front that separates these two traffic phases (see Kerner et al. 2013b).

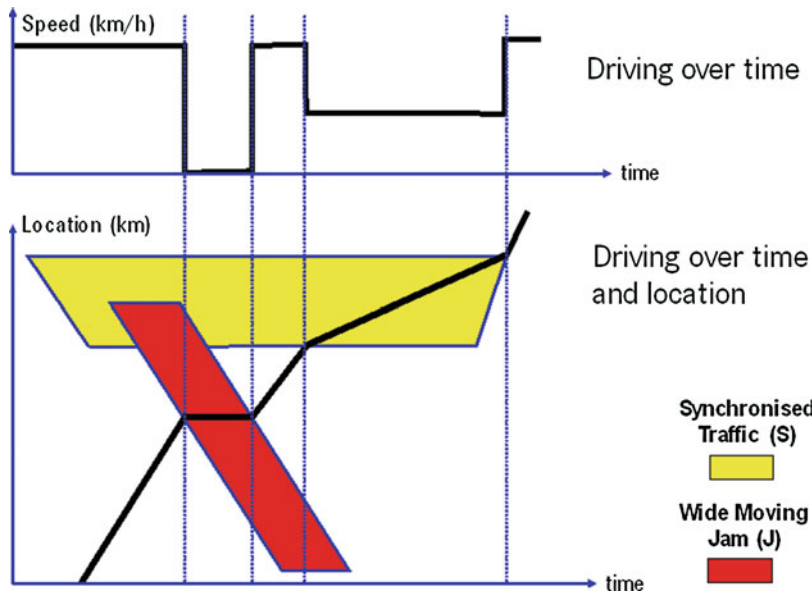
Traffic Prediction of Congested Patterns,

Fig. 5 Reconstructed trajectories from aerial photography by Treiterer (1975)



Traffic Prediction of Congested Patterns,

Fig. 6 Vehicle drives in traffic (abstract level) (see http://en.wikipedia.org/wiki/Three-phase_traffic_theory)



In particular, it is possible to distinguish the following points for transition points between different traffic phases along a vehicle trajectory (Fig. 7):

- From free flow to synchronized flow (denoted by F_S transition)
- From synchronized flow to free flow (S_F transition)
- From free flow to a wide moving jam (F_J transition)
- From synchronized flow to a wide moving jam (S_J transition)
- From a wide moving jam to free flow (J_F transition)
- From a wide moving jam to synchronized flow (J_S transition)

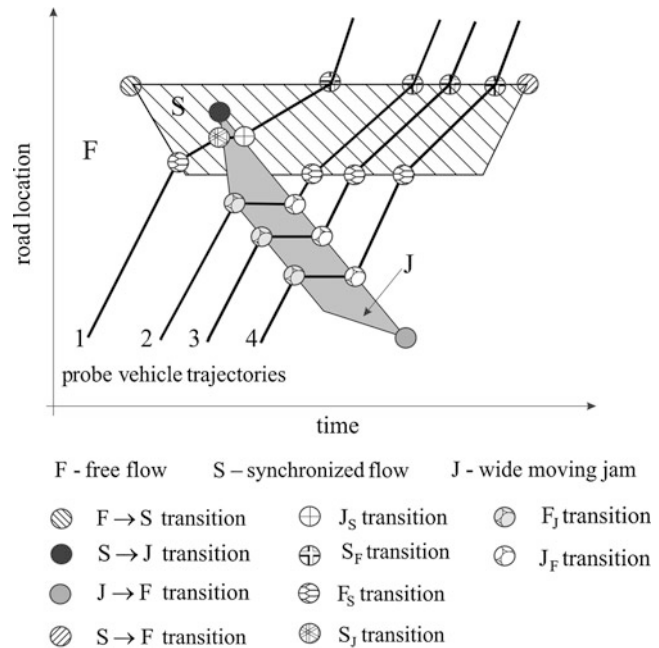
There is a qualitative difference between phase transitions in traffic and points for phase transitions along vehicle trajectories at which drivers observe a phase transition. While moving through a congested pattern, rather than an initial phase transition, a driver usually observes a transition point between different traffic phases upstream and downstream of the front separating these

traffic phases. This is qualitatively explained and shown in Fig. 7: an $F \rightarrow S$ transition, which leads to the emergence of synchronized flow and subsequent propagation of the front between two traffic phases F and S, has occurred earlier and at another road location than a location at which a driver of the probe vehicle 1 encounters the front between the phases F and S (transition point for FS transition along the trajectory 1 in Fig. 7). An $S \rightarrow J$ transition, which leads to the emergence of a wide moving jam within synchronized flow and subsequent propagation of the upstream front of a wide moving jam, i.e., the front between two traffic phases S and J, has also occurred earlier and at another road location than a location at which the driver encounters the jam (see, e.g., point for S_J transition on the trajectory 1 in Fig. 7).

Thus, transition points on a vehicle trajectory at which the abovementioned phase transitions (a)–(f) can be identified are usually not associated with the location and time instant at which phase transitions have initially occurred in traffic flow. For this reason, phase transitions are denoted by $F \rightarrow S$ transition, $S \rightarrow F$ transition, etc. In contrast, transition points for phase transitions along vehicle trajectories of probe vehicles are denoted by F_S

Traffic Prediction of Congested Patterns,

Fig. 7 Qualitative explanation of the difference between transition points for phase transitions along trajectories of four probe vehicles and phase transitions in traffic. *White regions: free flow (F); a dashed region, synchronized flow (S); gray region, wide moving jam (J)* (Taken from Kerner et al. 2013b)



transition, S_F transition, etc. (see items (a)–(f) above).

After successful matching of the GPS data of the vehicles, we can show space-time diagrams for different road stretches and the related GPS raw data. In the next examples, we always just simply color the GPS data according to the following legend: (i) red, 0–25 km/h; (ii) yellow, 25–60 km/h; and (iii) green, >60 km/h. We show examples from freeway segments in Germany and UK: in all cases based on the coloring qualitatively, the regions of the three traffic phases can be identified. The similarities of the congestions in different countries based on GPS raw data diagrams show similar results as in earlier investigations based on detector data (Rehborn et al. 2011).

The anonymized data is obtained from connected vehicles of a large fleet from real-world customers driving on the public road network. These vehicles send the GPS location alongside with a time stamp and session identifier to a secure backend to be then transferred to the traffic service provider (TSP), who in return delivers precise and latest traffic information such as traffic jams, dangerous end of jams, and temporary road closures to the connected vehicles. The data comprises the raw GPS location that after map matching represents a trajectory on the given road section. Depending on the vehicle model and device, the data has a sampling rate of 5–10s for the GPS points in each trajectory.

Figure 8 shows the freeway section A8 coming from “Kirchheim,” passing the city of Stuttgart

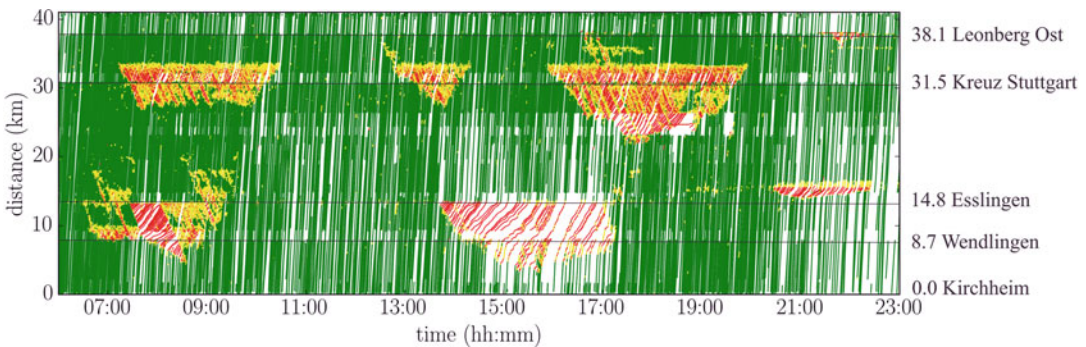
going toward “Pforzheim.” Marked on the side are bottlenecks that influence the traffic heavily. For once one can see the usual traffic patterns at peak hours from people coming (06:30–10:00 h morning peak) to work and leaving for home (16:00–20:00 h evening peak) at “Kreuz Stuttgart,” which is a major intersection on this freeway. At bottleneck “Esslingen” two large mega jams evolved after an accident.

In Fig. 9 the map-matched raw GPS data is shown for the M4 going toward London city center. As with the A8 section in Germany, the major jams are happening during the peak hours coming and going to work. Similar traffic patterns are observed on other freeways in other countries.

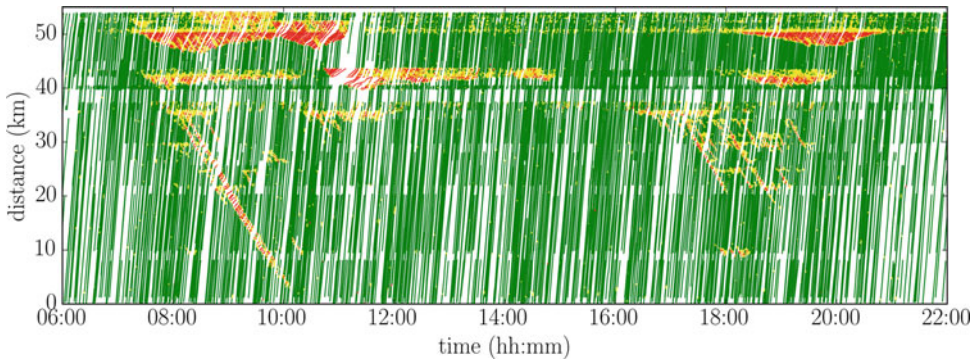
It is clear that with more data, the preciseness of the traffic reconstruction increases. The figures support a qualitative judgment that with approximately one GPS probe per minute (i.e., about 60 connected vehicles per hour which is about 1% penetration rate of the total traffic flow for a three-lane freeway with 6000 vehicles/h in peak times) on average, the quality is sufficient for a traffic service.

In more detail, we show a space-time diagram of GPS raw data over 4 h and more than 35 km of the freeway A5 (Figs. 10 and 11).

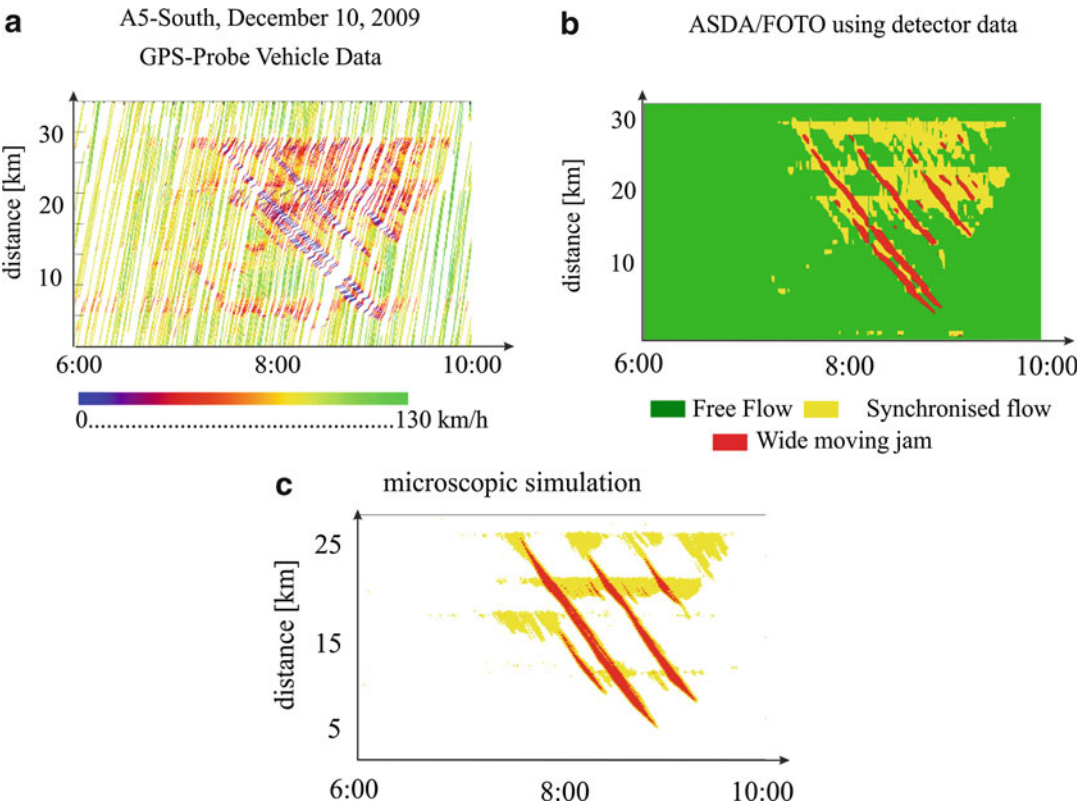
A general goal of a traffic information service would be the following request from customer perspective: “Show all congestions (hit rate) and do not show ones, which are not there (false alarms).” We use as one first approach a microscopic Kerner-Klenov traffic simulation (Kerner



Traffic Prediction of Congested Patterns, Fig. 8 Raw map-matched GPS data for 8 April 2016 for the A8 section. Horizontal black lines illustrate main locations of intersections



Traffic Prediction of Congested Patterns, Fig. 9 Raw map-matched GPS data for 5 April 2016 for the M4 section



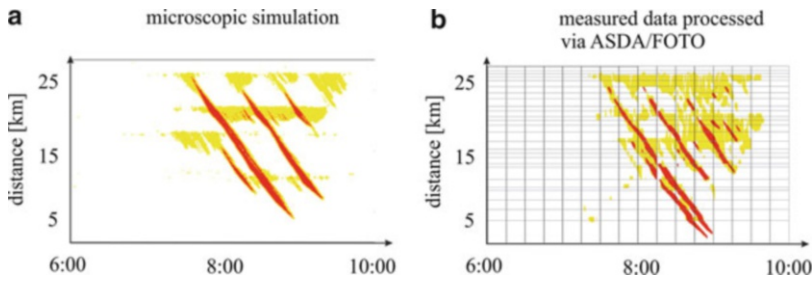
Traffic Prediction of Congested Patterns, Fig. 10 Overviews of GPS probe vehicle data (a) and road detector data (b) measured on 10 December 2009 on A5-South freeway in Germany within a congested pattern: (a) Vehicle trajectories in space and time. (b) The three

traffic phases reconstructed with FOTO and ASDA models from raw road detector data. (c) one realization of microscopic simulation for measured traffic situation in (b). (Taken from Kerner et al. 2013b)

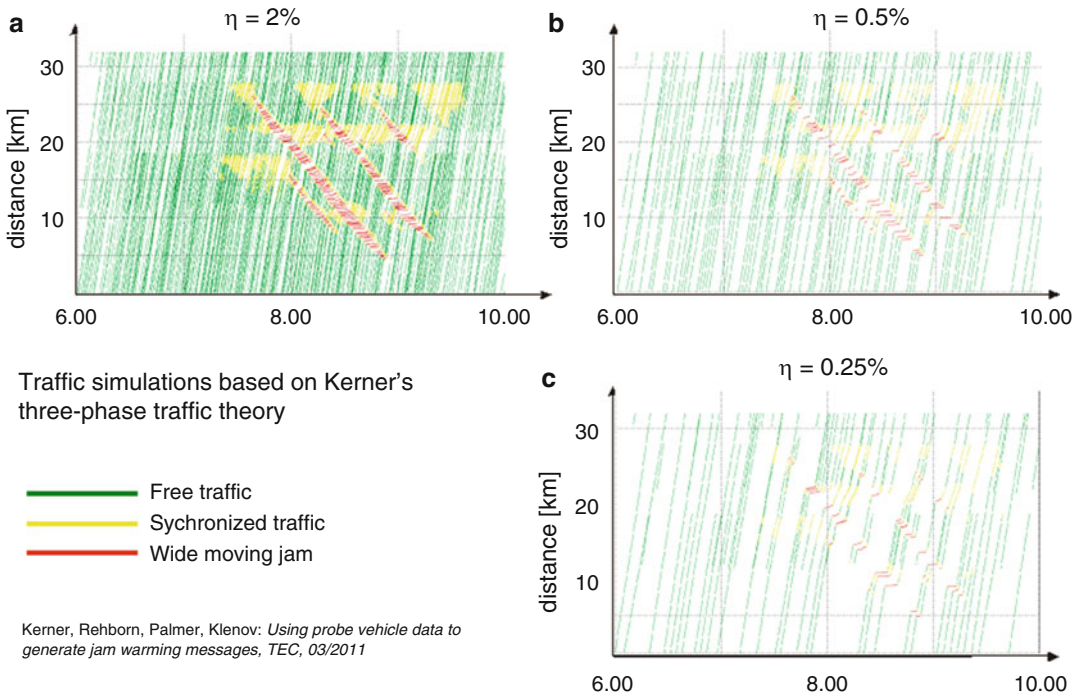
2004) to simulate a traffic pattern with congestions based on vehicle data only. Figure 12 (taken from Kerner et al. 2011) shows a traffic state recognition based on Kerner’s three-phase traffic

theory based on different percentages of sending (“equipped”) vehicles.

The propagating wide moving jams are clearly visible based on 2% FCD probes. With 0.5%



Traffic Prediction of Congested Patterns, Fig. 11 Comparison of realization 1 of microscopic simulations (a) with real measured data processed via the ASDA and FOTO methods (b) (Figure (b) is taken from Fig. 5(b). Taken from Kerner et al. 2013b)



Traffic Prediction of Congested Patterns, Fig. 12 Traffic state recognition based on different probe densities (Kerner et al. 2011)

(Fig. 12b) the propagating structure is still detectable, but the regions of synchronized flow are less precise.

Based on 0.25% of FCD percentage, even the propagating wide moving jams are becoming relatively imprecise: as a consequence of this approach, detailed structure of traffic congestions can be reconstructed with 2%, from 0.5% a travel (and delay) time estimation is feasible. Below that, the traffic service quality is decreasing to a

lower quality level and coming closer to random traffic message generation. Therefore, the number of FCD over space and time is one of the key issues for the traffic service quality.

Some details of the traffic reconstruction using GPS data are described in Kerner et al. (2013b): the regions of the congested phases can be reconstructed based on sufficient GPS data. For the traffic service the reconstructed congested regions are translated into traffic message

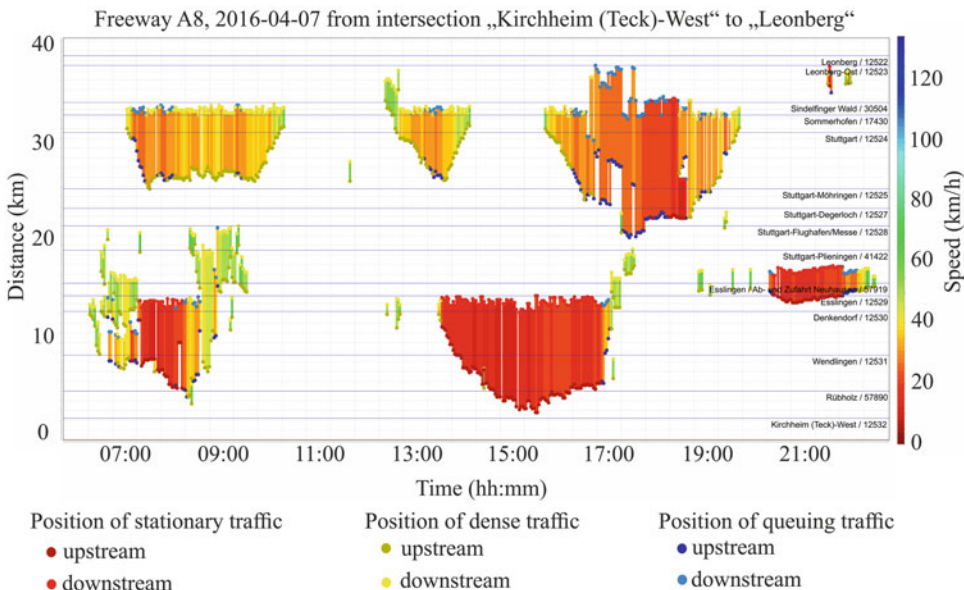
protocols: TMC (Traffic Message Channel, https://en.wikipedia.org/wiki/Traffic_message_channel) and TPEG (Transport Protocol Expert Group, <https://en.wikipedia.org/wiki/TPEG>) are two technical options to transfer traffic information from a traffic center to a navigation system in a vehicle. Figure 13 shows the traffic messages in the TMC protocol using the raw GPS data shown in Fig. 8: the colored lines indicate the average vehicle speeds within the traffic congestions with lower speeds with more and darker red. The length of these lines corresponds to the length of each related TMC traffic message. The smaller bullets indicate the beginning and ending of the upstream and downstream position of the related TMC event messages: (i) red indicates the stationary traffic, (ii) yellow indicates the dense traffic message, and (iii) blue indicates queueing traffic events. From customer perspective, the navigation system shows the traffic message with both event information (stationary, queuing, or dense traffic) and each length of the congestion: this is in very good correlation to the GPS raw data (Fig. 8), and each user has received precise traffic information in his system. The TMC

traffic message averages the traffic phases to one length; therefore, for an individual driver on a specific lane, the speed might differ a little bit. Overall, based on the TMC message, the delay and arrival times can be calculated in the navigation systems.

Beyond the current traffic situation predicted information would be a next step to enhance the customer quality: algorithm and technologies are mentioned in this contribution, and with the increase of GPS data and its archiving, the prediction capabilities will become more and more ready for market introduction: a customer route is in principle a trip into the future, and, therefore, a predicted traffic information is of very high usefulness for navigation systems.

Prediction of Upstream Fronts of Traffic Phases (In Particular “Jam Warning”) Based on Probe Vehicle Data

The availability of masses of probe vehicle data allow the prediction of the upstream front of the wide moving jam phase: this is very important for a jam front warning for a vehicle which



Traffic Prediction of Congested Patterns, Fig. 13 TMC traffic messages over space and time corresponding to the raw data of Fig. 8. Horizontal blue lines illustrate locations of intersections

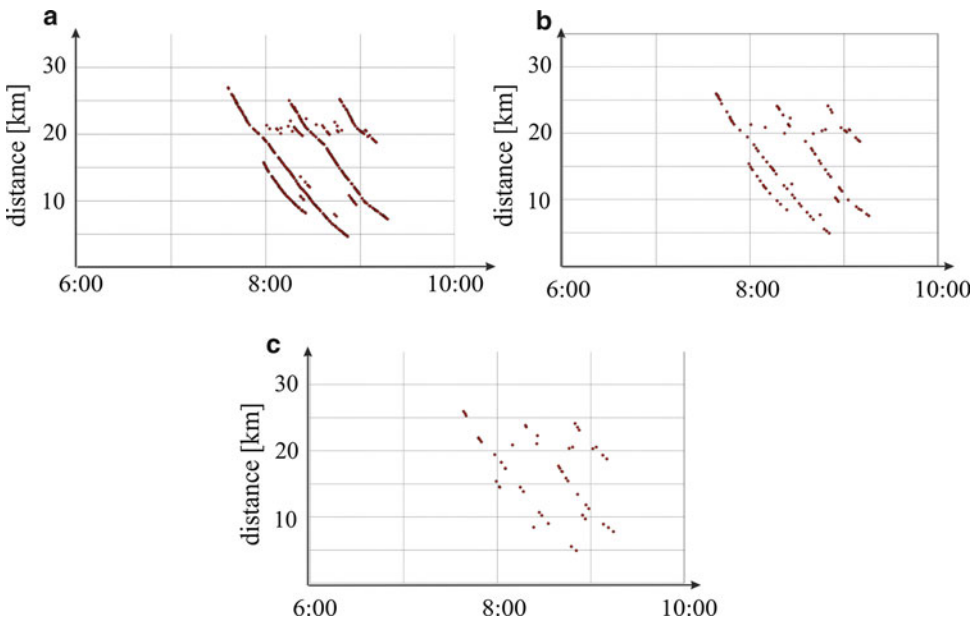
approaches such a front driving currently in free flow. The traffic phase identification along a probe vehicle trajectory explained above allows us to generate jam warning messages as follows: (i) a jam warning message can be generated by a probe vehicle when either a F_J transition (transition from free flow to wide moving jam) or an S_J transition (transition from synchronized flow to wide moving jam) is identified along the vehicle trajectory. Then a sequence of these jam warning messages generated by all probe vehicles (Fig. 14) allows us to distinguish a spatiotemporal behavior of the propagation of the jam upstream front on the road. The upstream jam front propagation can be used for the warning of all other vehicles moving upstream of the jam front. Figure 14 illustrates an example of all F_J transition points with different percentages of probe vehicles. Similarly, this can be done for each of the phase transitions (see Kerner et al. 2013b).

Through the use of simulated trajectories of probe vehicles randomly distributed in traffic flow in Fig. 12, we can find the dependence of

jam warning messages on penetration rate of probe vehicle data (Fig. 14). The propagation of the front of the transition recognized by the probe vehicles can also be used for the warning of all other vehicles moving upstream of the wide moving jam front. The model of traffic phase identification allows us also to generate messages indicating the possibility of vehicle acceleration from a wide moving jam to either synchronized flow (J_S transition) or to free flow (J_F transition) (Kerner et al. 2013b). The propagation of the associated front of vehicle escaping from the jam found through the probe vehicles can help other vehicles estimate the time to acceleration from the jam.

Reconstruction of Synchronized Flow Patterns at Traffic Signal in City Traffic with the Use of Probe Vehicle Data

In contrast to freeway traffic, traffic flow in city traffic is regularly interrupted by traffic signals (e.g., Dion et al. 2004; Gartner and Stamatiadis



Traffic Prediction of Congested Patterns, Fig. 14 Detection of upstream jam fronts (“jam warning messages”) with different percentage of simulated probe

vehicles: (a–c) percentage of probe vehicles $\eta = 2\%$ (a), $\eta = 0.5\%$ (b), and $\eta = 0.25\%$ (c). Traffic situation from Fig. 12 (Taken from Kerner et al. 2013b)

2008; Geroliminis and Skabardonis 2011; Newell 1965; Robertson 1969; Webster 1958). The two main types of city traffic at a traffic signal are undersaturated traffic and oversaturated traffic. In undersaturated city traffic, all vehicles waiting in a queue at a red traffic signal can pass the signal location during the next green phase. In the opposite case, city traffic is characterized as oversaturated: some of the vehicles in the queue cannot pass the signal during the next green phase, resulting in growth of the congested area in upstream direction.

In classical theory of city traffic (see Dion et al. 2004; Gartner and Stamatiadis 2008; Geroliminis and Skabardonis 2011; Newell 1965; Robertson 1969; Webster 1958), oversaturated city traffic is assumed to consist of moving queues that propagate upstream and in which vehicles stop, and regions between the moving queues in which vehicles move from one moving queue to the next downstream moving queue (Fig. 15a).

However, Kerner et al. concluded from numerical simulations with a stochastic Kerner-Klenov

model (Kerner 2004, 2009; Kerner and Klenov 2002, 2003) that synchronized flow should exist in oversaturated city traffic as well (Kerner et al. 2013a). In synchronized flow patterns, vehicles move with a speed considerably lower than free flow speed and, in contrast to moving queues, do not stop (Fig. 15b).

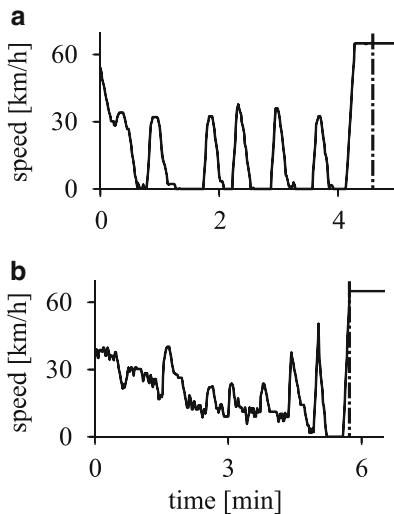
To study empirical spatiotemporal traffic patterns in oversaturated city traffic, we use anonymized GPS probe data from navigation devices. A navigation device measures the GPS locations of a vehicle at time instants with a fixed time interval, e.g., often 5 or 10 s. Usually, every 2 min the recorded GPS locations are sent to a central server. The data collected on the server allow the reconstruction of vehicle trajectories and speed profiles and the calculation of macroscopic parameters such as the mean speed along a trajectory.

We investigate probe vehicle data from two sections of urban main roads in the city of Düsseldorf, Germany: “Völklinger Straße” and “Südring” (see Fig. 16a, b). On these road sections, traffic breakdowns resulting in oversaturated city traffic are often observed, as is confirmed by stationary detectors (labeled “DV” and “DS” in Fig. 16a, b) that deliver the flow rate and average speed for each minute (Kerner et al. 2014).

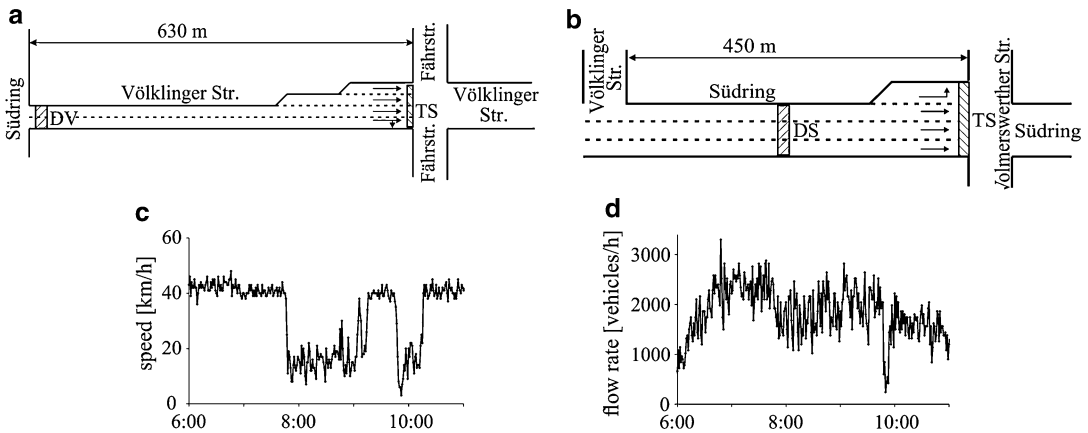
Oversaturated city traffic results from urban traffic breakdowns that are characterized by the following phenomena observed in the speed and flow rate measured by stationary detectors (Fig. 16c, d) (Kerner et al. 2014). At a small flow rate, there is free flow. When the flow rate increases considerably, a drop in the speed is observed. There are two qualitatively different cases:

- (i) During the drop in speed, the decrease in the flow rate is considerably smaller than the drop in speed. This case is called “speed breakdown” (7:48–9:15 h in Fig. 16c, d).
- (ii) Both the speed and the flow rate drop considerably. This case is called “speed and flow breakdown” (9:47–9:54 h in Fig. 16c, d).

These two cases of breakdown result in qualitatively different patterns of oversaturated city



Traffic Prediction of Congested Patterns, Fig. 15 Single-vehicle speed profiles from traffic flow simulations of oversaturated city traffic at a traffic signal: (a) sequence of moving queues of the classical theory (Dion et al. 2004; Gartner and Stamatiadis 2008; Geroliminis and Skabardonis 2011; Newell 1965; Robertson 1969; Webster 1958); (b) synchronized flow pattern. Dash-dotted lines indicate time instant when the vehicle passes the traffic signal (Taken from Kerner et al. 2013a)



Traffic Prediction of Congested Patterns, Fig. 16 (a, b) Schemes of road sections of “Völklinger Straße” and “Südring” (TS: traffic signal, “DV” and “DS” are stationary detectors). The speed limit is 60 km/h on both sections.

(c, d) City traffic detector data for speed and flow rate on 4 April 2013 for speed and flow rate measured at detector “DV” (Taken from Kerner et al. 2014)

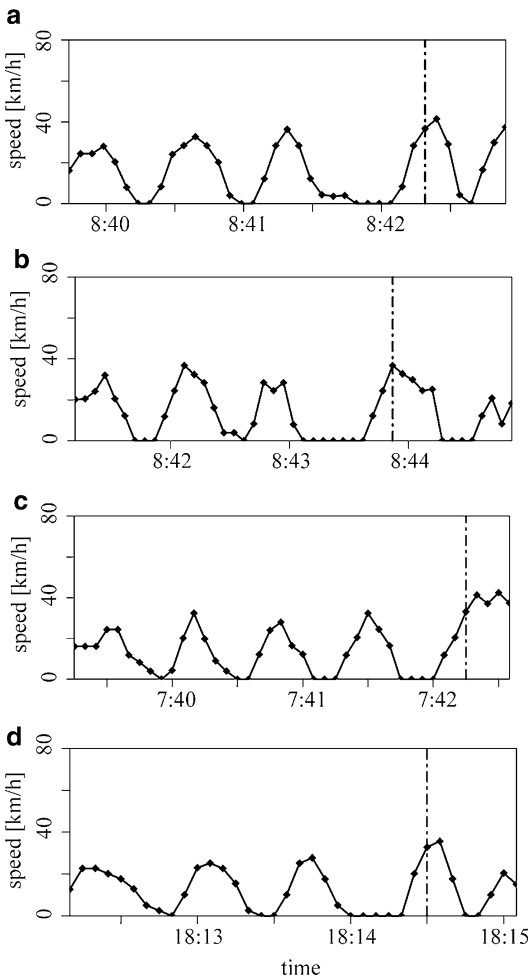
traffic: In case (i) of speed breakdown only, both moving queues and synchronized flow patterns are observed. On the other hand, in case (ii) of speed and flow breakdown at the detector (e.g., at detector “DV” shortly before 10:00 h), which results from heavy bottlenecks due to, e.g., accidents or roadworks, a mega-queue is observed (Kerner 2009). Synchronized flow does not occur in case (ii) of speed and flow breakdown. Although we present and study real field GPS probe data from only two different road sections sketched in Fig. 16a, b all results presented below are common ones for many other streets of different cities in Germany studied in this paper.

In empirical GPS probe data from these two different road sections in the City of Düsseldorf, we identified the two basic patterns of over-saturated city traffic resulting from speed breakdown: sequences of moving queues (Fig. 17) and synchronized flow patterns (Fig. 18). There are apparent qualitative differences between these two types of traffic patterns: In the case of moving queues, the drivers accelerate from speed zero several times when leaving each of the moving queues. On the other hand, in the examples depicted in Fig. 18, the vehicles stop only once at the signal location. Upstream of the location of these stops,

the vehicles traverse synchronized flow patterns. In these synchronized patterns, the vehicles do not stop, and thus there is no acceleration from zero.

Fuel Consumption in Road Networks

We studied empirical microscopic features of synchronized flow in congested traffic measured on freeways and in urban networks. The microscopic empirical data is measured in probe vehicles (single-vehicle data) moving on freeways and city traffic. Additionally to traffic characteristics (vehicle speed as function of time and vehicle coordinates), some of this data include fuel consumption data measured in vehicles. Based on these measurements, we found a relation between traffic phases and fuel consumption. It turned out that fuel consumption in congested traffic depends crucially on traffic phases in which traffic currently is: even at same average speed, fuel consumption in synchronized flow is considerably smaller than in sequence of moving jams (stop and go traffic). We have found the decrease in consumption in synchronized flow of city traffic in comparison with moving queues. The three-phase traffic theory can explain these empirical microscopic results.

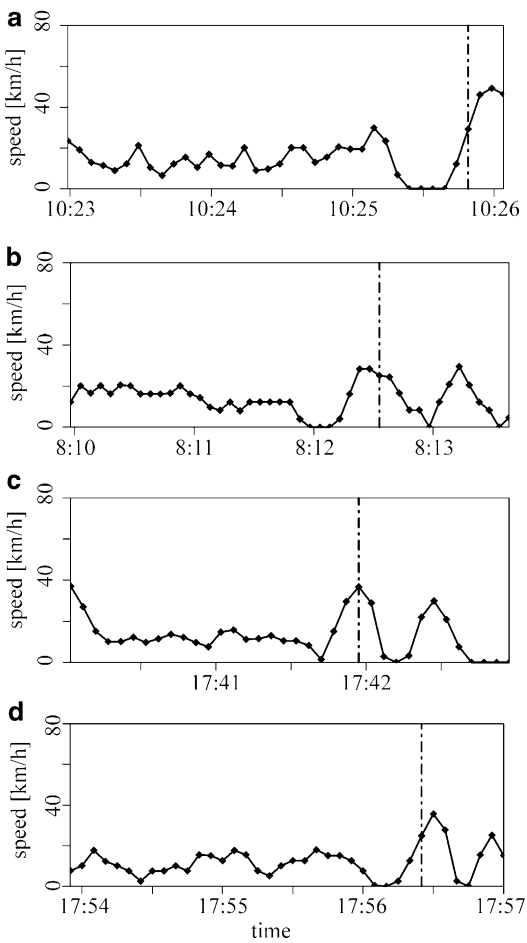


Traffic Prediction of Congested Patterns, Fig. 17 Single-vehicle speeds from GPS floating car data of vehicles traversing sequences of moving queues on “Völklinger Straße” (a–c) and “Südring” (d). *Dash-dotted lines* indicate the first recorded time instant when the vehicle has passed the traffic signal (Taken from Hemmerle et al. 2016a)

Empirical Microscopic Fuel Consumption Data

A set of empirical measurements of vehicles driving on the public freeway A5 near Frankfurt in Germany is illustrated. As example, 1 h of empirical vehicle trajectories of this data set is shown in Fig. 19. Each line represents a vehicle which is driving over the length of 17 km of the considered road stretch.

The first vehicle in Fig. 19 starts in free traffic flow till it has to reduce its velocity considerable at



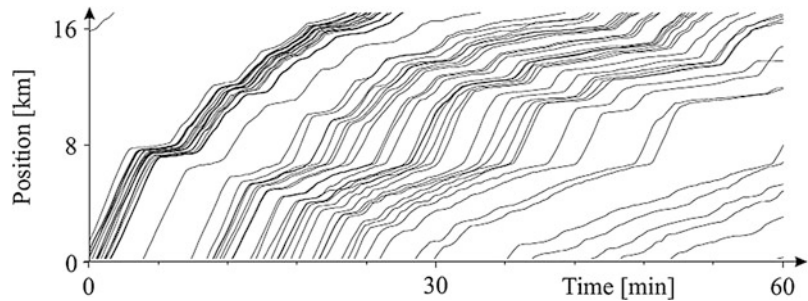
Traffic Prediction of Congested Patterns, Fig. 18 Single-vehicle speeds from GPS floating car data of vehicles traversing synchronized patterns on “Völklinger Straße” (a, b) and “Südring” (c, d). *Dash-dotted lines* indicate the first recorded time instant when the vehicle has passed the traffic signal (Taken from Hemmerle et al. 2016a)

8 km caused by a traffic state transition from free flow to wide moving jam. Together with the other vehicles, the three traffic phases (Kerner 2004, 2009) (free traffic flow, synchronized flow, and wide moving jam) can be reconstructed using probe vehicle data (Kerner et al. 2011).

Within our data set each of the three phases occurs. The velocity is one indication of which traffic phase a vehicle is currently driving through. Examples of typical velocities recorded in our data set of free flow visualized in Fig. 20a,

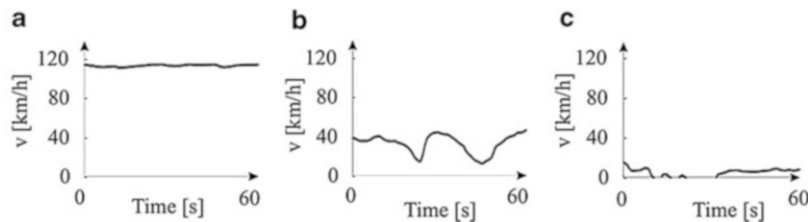
Traffic Prediction of Congested Patterns,

Fig. 19 Empirical trajectories of vehicles driving through a congested traffic pattern on a freeway (Koller et al. 2016)



Traffic Prediction of Congested Patterns,

Fig. 20 Empirical microscopic vehicle data in (a) free traffic flow, (b) synchronized flow, and (c) wide moving jam



Traffic Prediction of Congested Patterns, Table 1 Empirical measurements of one vehicle during 1 s

Time	v (km/h)	a (m/s ²)	W _{cum} (m)	C _{cum} (l)	c _t (l/h)	c _m (l/100 km)
17:45:15.054	—	0.0	160,909	—	—	—
17:45:15.147	79.08	−0.098	160,911	14.227	7.405	8.823
17:45:15.241	—	0.0	160,913	—	—	—
17:45:15.335	79.03	−0.098	160,915	14.227	7.405	8.823
17:45:15.444	—	−0.098	160,917	14.227	7.405	4.426
17:45:15.538	78.97	0.0	160,920	14.227	7.405	4.426
17:45:15.632	78.91	—	—	—	—	—
17:45:15.741	78.86	−0.196	160,924	14.228	7.405	4.426
17:45:15.835	—	0.0	160,926	14.228	3.855	4.426
17:45:15.960	78.91	—	160,928	14.228	3.855	4.426

synchronized flow in Fig. 20b and wide moving jam in Fig. 20c.

We examine empirical data read from the CAN (Controller Area Network) serial bus system of vehicles driving on German freeways. Instantaneous velocity, acceleration, and consumption were recorded in parallel several times per second. To gain an impression, a sample of the data recorded in one vehicle during 1 s with the serial CAN bus system is shown in Table 1.

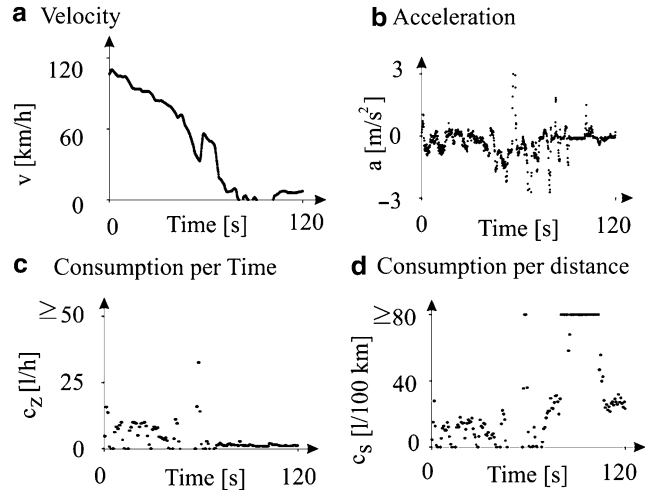
The first column of Table 1 shows a time stamp of the measurement in the format hour, minute, second, and milliseconds. The minimum temporal distance between two measurements is approximately

100 ms. For some time stamps, values of some variables are not available. During this example, the vehicle's velocity v is slightly reducing which results in negative values in the measured acceleration a . The odometer W_{cum} of the vehicle indicates that it has passed a stretch with a length of 19 m and the cumulated fuel consumption C_{cum} increased within that second by 0.001 l. The time-related consumption c_z and milage-based consumption c_m were recorded several time, but the value is updated only once per second.

Detailed empirical measurements of one vehicle during a traffic state transition from free traffic flow to wide moving jam are shown in Fig. 21.

Traffic Prediction of Congested Patterns,

Fig. 21 Empirical vehicle trajectory during a traffic state transition from free traffic flow to wide moving jam



The velocity (Fig. 21a) decreases from over 100 to 0 km/h. Beside strong decelerations up to 3 m/s^2 also acceleration up to 3 m/s^2 were measured (Fig. 21b). The consumption was determined per time (Fig. 21c) and per distance (Fig. 21d). During the deceleration, the vehicle consumptions per time are in average higher than during the vehicle standstill in the second half of the diagram (Fig. 21c). Reversed to that the fuel consumption per distance is low during the deceleration and goes up to infinity if the vehicle standstill (Fig. 21d).

In total, we examine vehicle trajectories with a length of approximately 5444 km, 5.6 million velocity and 3.6 million consumption measurements. The empirical velocity data is used to reconstruct the traffic phases by applying the in-vehicle traffic state detection (Kerner et al. 2011) using probe vehicle data. After that, the consumption data can be assigned to the three traffic phases. The vehicles drove 130 h in free traffic flow, 52 h in the phase of synchronized flow, and 13 h in the phase of wide moving jam. The distribution of all consumption measurements per traffic phase is shown in Fig. 22 as boxplot. This means that the black box covers 50 percent of all measurements and the anchors indicate the minimum and maximum consumption value.

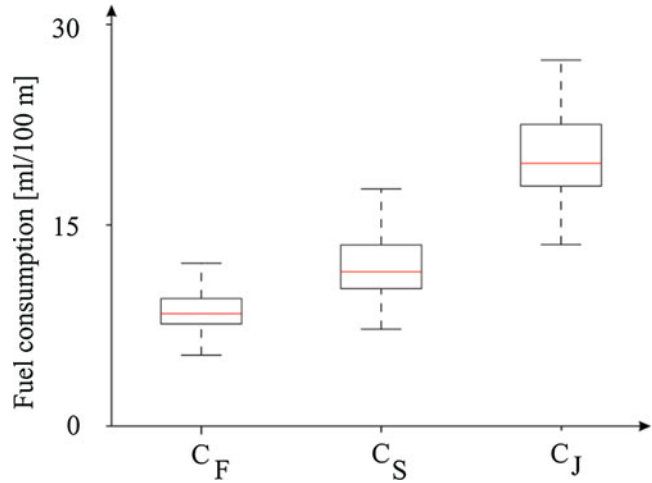
Empirical Microscopic Probe Data from Navigation Devices

To study microscopic empirical spatiotemporal features of oversaturated city traffic, we use anonymized GPS (global positioning data) probe data from navigation devices (PND: personal navigation device). A navigation device of this kind measures and stores the locations x_i of a vehicle at associated time instants t_i , where $\Delta t = t_i - t_{i-1}$ is a fixed time interval. For the data used in this article, depending on the type of PND, $\Delta t = 5 \text{ s}$ or, occasionally, 10 s . Every 120 s, the measured data are uploaded to a server. From a complete sequence of positions x_i and associated time instants t_i of a vehicle on a road section, its trajectory in space and time can be reconstructed. In addition, the mean speeds v_i between two subsequent GPS measurements can be derived, which provides a good approximation of the vehicle speed along the trajectory. On multilane roads, the lane a vehicle is moving on cannot be determined by means of GPS data from PNDs.

This data set of empirical microscopic probe data from navigation devices does not contain any data about fuel consumption. To estimate the fuel consumption of these empirical trajectories, we used a microscopic empirical fuel consumption matrix which is explained in the following section.

Traffic Prediction of Congested Patterns,

Fig. 22 Empirical consumption distribution per traffic phase for freeways: C_F free traffic flow, C_S synchronized flow, and C_J wide moving jam (Koller et al. 2016)



Physical Features of Traffic Phases in Congested Traffic and Fuel Consumption

Empirical measurements on freeways show that fuel consumption in spatiotemporal areas of synchronized traffic and wide moving jams are significantly higher than fuel consumption in free traffic flow. Driving through an area of synchronized flow, we observed an average increase of the fuel consumption by 1.42 and 2.34 in wide moving jam.

The analyses of several empirical vehicle trajectories in urban areas with traffic signals show an increased fuel consumption during synchronized flow of 2.0 and 2.9 while passing wide moving queues in comparison to fuel consumption while the traffic signal is in the state of undersaturation.

Therefore, the influence of congested traffic to fuel consumption in oversaturated city traffic is larger than in congested traffic on freeways. A typical wide moving jam has a spatial expansion of 1023 m and increases the fuel consumption of a single vehicle by 119 ml.

Consider vehicle fuel consumption and travel time depending on the traffic phase of congested traffic can be relevant for practical application due to the following reasons.

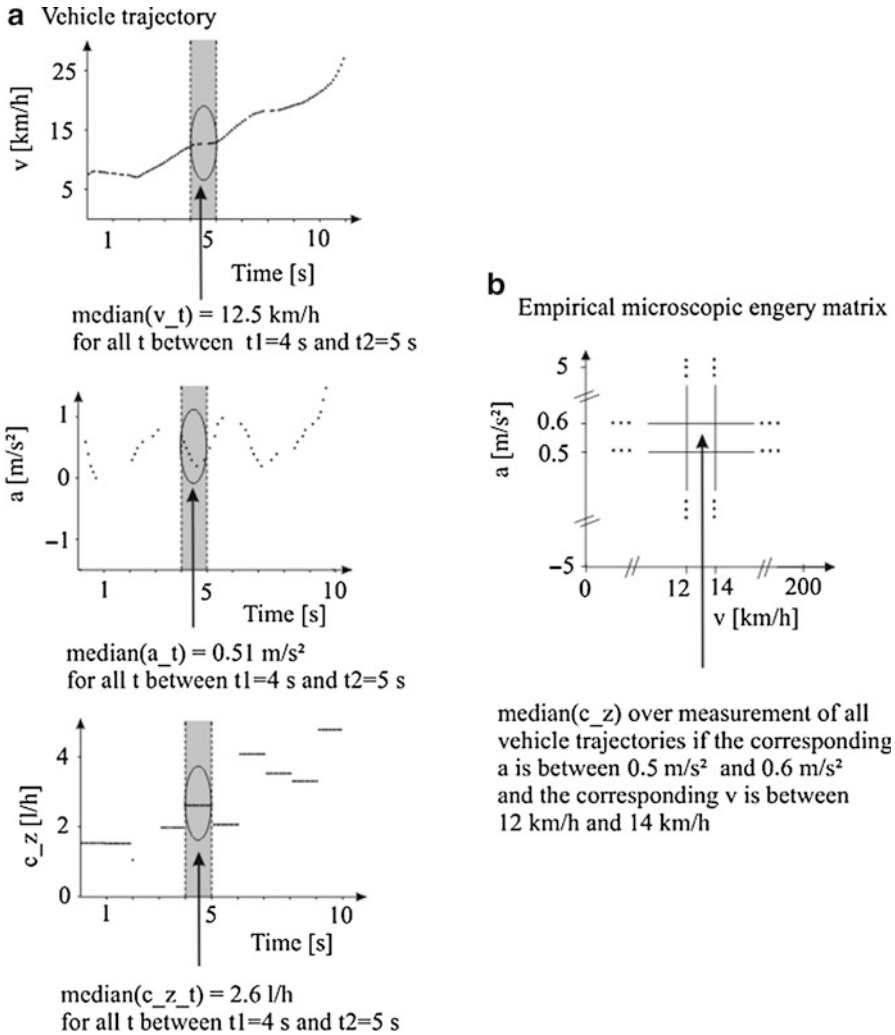
Vehicle fuel consumption in congested traffic is higher than in free flow. Fuel consumption in synchronized flow is lower than fuel consumption in wide moving jam. If the traffic volume is high, a

breakdown in the traffic network cannot be prevented. However, applying intelligent distribution of traffic synchronized flow can be tolerated, but emerging wide moving jams can be countered. Despite of congested traffic, the fuel consumption in the whole network can be reduced.

Empirical Microscopic Energy Matrices

For energy efficiency calculations, we use empirical microscopic consumption matrices. The empirical microscopic matrices were aggregated by grouping energy values into matrix elements according to their associated speed and acceleration values. For each matrix element, the energy median was then calculated (see Fig. 23). Neither data from test bench measurements nor consumption models were used.

An empirical microscopic fuel consumption matrix for a conventional medium-sized vehicle with a combustion engine is shown in Fig. 24a which shows normalized values between 0 and 1. Empirical field trial data of this vehicle measured over 90 h on 20 different days were used. The vehicle was driven by different drivers on different roads, both freeways and urban roads. More than ten million data points were recorded and aggregated. The aggregation intervals are 2 km/h and 0.1 m/s², respectively. In addition, measurements from vehicle with combustion engine we also analyzed electrically driven vehicles. An empirical microscopic matrix for an



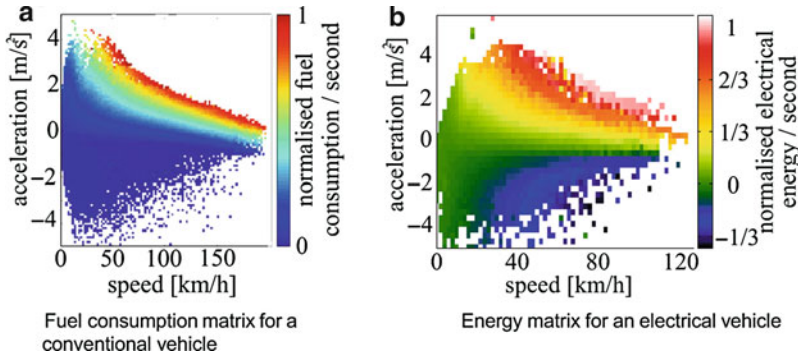
Traffic Prediction of Congested Patterns, Fig. 23 Creation of an empirical microscopic energy matrix through aggregation of empirical measurements

electrical vehicle is shown in Fig. 24b. Due to energy regeneration associated with braking, several energy values are negative if the acceleration of the vehicle is negative too. The empirical microscopic matrix for an electrical vehicle was aggregated on the basis of about 146,000 data points. In the electric vehicle, no acceleration signal was recorded in the field trial. The acceleration values used here were therefore derived from the speed. The resolution of the measured velocity is 1 km/h. The aggregation intervals are 2 km/h (as in the case of the conventional vehicle) and 0.278 m/s^2 (resulting from the measurement

accuracy of the velocity values from which the acceleration was derived).

Cumulated Acceleration and Energy Efficiency of Vehicles

Empirical microscopic fuel consumption matrices cannot directly be used for online assessment of fuel consumption, because on the traffic server side, the current velocity, acceleration, and consumption values are not available for each second and vehicle on any considered road stretch. To assess the dependency of traffic patterns on energy efficiency, we use a parameter that can



Traffic Prediction of Congested Patterns, Fig. 24 Empirical microscopic energy matrices: (a) Empirical microscopic fuel consumption matrix of a conventional medium-sized vehicle with a combustion engine which drives over 83 h (ten million data points) in various

traffic situations in the real road network (Koller et al. 2014), (b) empirical microscopic energy consumption matrix for an electrical vehicle containing about 146,000 data points (Hemmerle et al. 2016c)

feasibly be calculated on the basis of position data from navigation devices and that correlates with the energy efficiency of vehicles. In particular, taking into account the aforementioned dependency of energy efficiency on acceleration, we choose a macroscopic parameter which sums up how often and how strongly a vehicle accelerates along a road section. The best information on these acceleration processes available from position data is the speed differences $v_{n+1} - v_n$ between two successive measurements in a vehicle traveling on a road. The cumulated vehicle acceleration per road length (or “cumulated acceleration” for short), which has been introduced by Kerner for a reliable online assessment of fuel consumption in vehicles, is a sum of positive speed differences between position measurements (Kerner 2014). The cumulated acceleration A per length, which refers to a vehicle traveling on a road section of length L , is defined by the formula (from Kerner 2014):

$$A = L^{-1} \sum_{n=1}^{N-1} (v_{n+1} - v_n) \Theta[v_{n+1} - v_n - \Delta v], \quad (1)$$

where $\Theta[x]$ is the Heaviside function, and $\Delta v \geq 0$ is used to reduce the effect of the error resulting from speed calculations from position data. This formula can be applied to any road stretch for

which vehicle trajectories are available even if the position data is available less than every second. The cumulated vehicle acceleration can be applied for any link of a traffic network, independent of the length of the link. Furthermore, it can be applied for any time interval Δt between successive position measurements.

The regenerative brake of an electrical vehicle allows energy recovery. To account for the energy recovery of an electrical vehicle, we define the cumulated deceleration D analogously to the abovementioned formula for the cumulated acceleration A (from Hemmerle et al. 2016c):

$$D = L^{-1} \sum_{n=1}^{N-1} (v_{n+1} - v_n) \Theta[-(v_{n+1} - v_n) - \Delta v]. \quad (2)$$

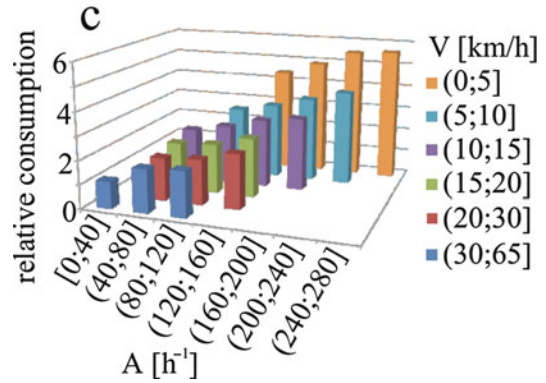
We used speed and acceleration time series of more than 400,000 simulated vehicle trajectories provided by Hermanns that reproduce various scenarios of city traffic (Hermanns et al. 2015, 2016). All simulated speed and acceleration time series were fed to the empirical microscopic consumption matrices (see Fig. 24). The simulated vehicle trajectories are classified according to their mean speed V along the trajectory and their cumulated acceleration A . For all classes with at least 50 vehicle trajectories, the average consumption was calculated. Macroscopic consumption

matrix for a conventional vehicle with a combustion engine (see Fig. 25) expresses the following dependency between mean speed, cumulated acceleration, and fuel consumption (Hemmerle et al. 2016a): (i) For mean speeds between 0 and 65 km/h, fuel consumption increases with decreasing mean speed, and (ii) fuel consumption increases with increasing cumulated acceleration.

For electrical vehicles, we chose a modified approach to express the energy balance (the difference between the energy consumed and the energy regenerated while driving) as a function of mean speed V , cumulated acceleration A , and cumulated deceleration D . To this end, we split the empirical microscopic energy matrix (Fig. 25) into two parts: one part that contains all combinations of speed and acceleration with nonnegative acceleration values and one part that contains all combinations with negative acceleration values. Both partial microscopic energy matrices were applied to the same simulated vehicle trajectories as were used for the macroscopic matrix for a conventional vehicle. As a result, two macroscopic matrices were calculated that are shown in Fig. 26. It expresses the dependency of electrical energy on mean speed V and cumulated vehicle acceleration A (Fig. 26a) as well as vehicle deceleration D (Fig. 26b) (Hemmerle et al. 2016c).

To discuss a comparison between the fastest route and the most energy-efficient route, hypothetical road network is introduced (see Fig. 27) within drivers who can choose either route ACB or route ADB to get from point A to point B (Hemmerle et al. 2016a). On both routes, oversaturated traffic occurs which causes the same mean speed V . Therefore, the shorter route ADB is the faster route. However, moving queues and synchronized flow patterns led to a much higher cumulated acceleration A for the route ADB. Based on these acceleration numbers, the energy consumption read from the macroscopic consumption matrix C multiplied with the route length for ACB is 5 and 5.86 for route ADB which makes ACB to the most energy-efficient route.

To discuss a comparison between most energy-efficient route for conventional vehicles and electrical vehicles, another hypothetical road network is introduced (see Fig. 28) within drivers can

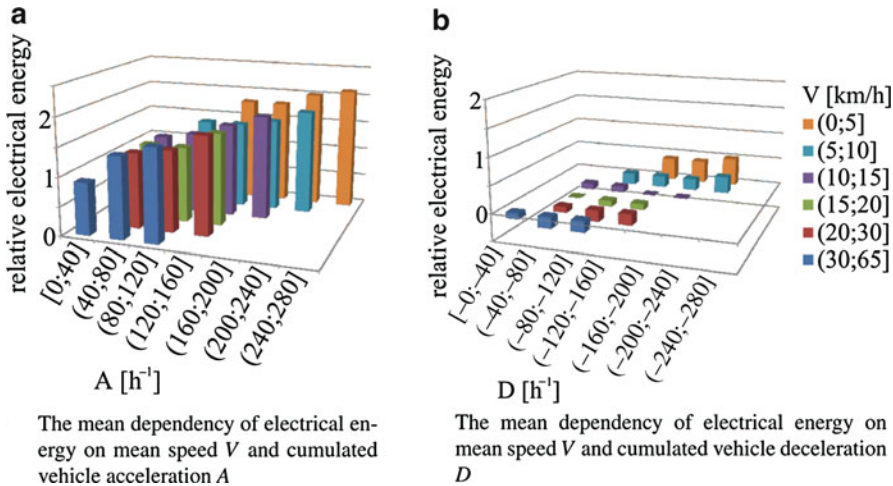


Traffic Prediction of Congested Patterns, Fig. 25 Macroscopic fuel consumption data based on their cumulated acceleration (Hemmerle et al. 2016b, c)

choose either route EF or route EGF to get from point E to point F (Hemmerle et al. 2016c). According to the length and the mean speed, the travel time for route EF with 300 s is much higher than the travel time for route EGF with 180 s as well as the cumulated acceleration A and cumulated deceleration D . The relative fuel consumption read from the macroscopic consumption matrix C (Fig. 25) multiplied with the route length for EF is 3.1 and 2.75 for route EGF. For electrical vehicles, the relative electrical energy E read from the macroscopic energy matrices (Fig. 26) multiplied with the route length for EF is 1.8 and 0.8 for EGF. While EGF is more energy efficient for a conventional vehicle, the congested route EF is more energy efficient for an electrical vehicle. This result indicates that the impact of oversaturated traffic on energy efficiency is stronger for a conventional vehicle than for an electrical vehicle.

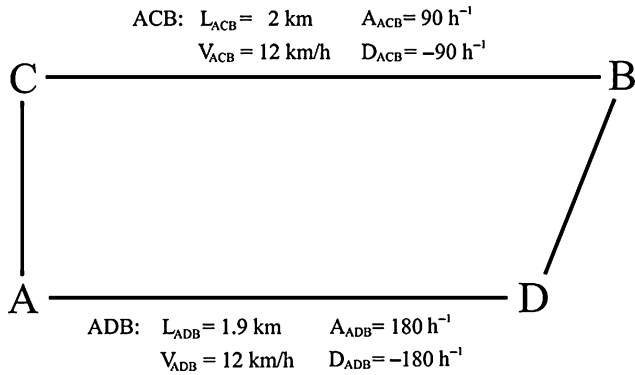
According to the energy consumption we conclude the following:

- (i) The cumulated acceleration is considerably greater for moving queues than for synchronized flow patterns.
- (ii) The energy consumption of vehicles increases with increasing cumulated acceleration.
- (iii) The first two items mean that the energy consumption of vehicles traversing moving queues is considerably greater than the energy consumption of vehicles traversing synchronized flow patterns.

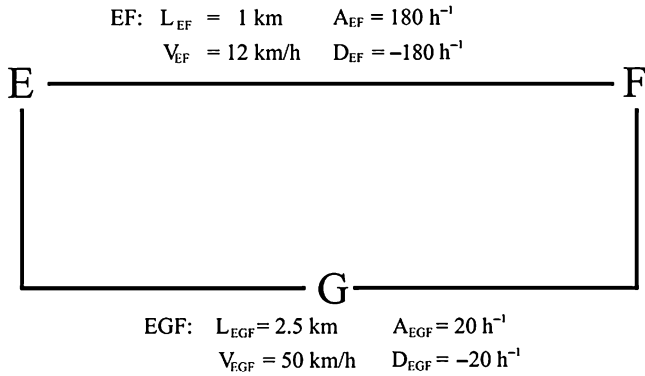


Traffic Prediction of Congested Patterns, Fig. 26 Macroscopic energy matrices for an electrical vehicle (Hemmerle et al. 2016c)

Traffic Prediction of Congested Patterns, Fig. 27 Schematic sketch of a subset of a hypothetical urban network with two alternatives routes ACB and ADB from point A to B (Hemmerle et al. 2016c)



Traffic Prediction of Congested Patterns, Fig. 28 Schematic sketch of a subset of a hypothetical urban network with two alternatives routes EG and EGF from point E to F (Hemmerle et al. 2016c)



- (iv) The impact of oversaturated city traffic on the energy efficiency of vehicles is stronger for a conventional vehicle than for an electrical vehicle.
- (v) In an urban road network, the fastest route between two points can be different from the most energy-efficient route.
- (vi) In an urban road network, the most energy-efficient route for a conventional vehicle with a combustion engine can be different from the most energy-efficient route for an electrical vehicle.

Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data

FOTO and ASDA Models

One of the engineering applications of this three-phase traffic theory are the FOTO and ASDA models proposed by Kerner and further developed by Rehborn, Aleksic, and Haug for recognition and tracking of spatiotemporal congested traffic patterns occurring at freeway bottlenecks (Kerner 1999b, 2004, 2008, 2009; Kerner et al. 1998, 1999, 2004; Kerner and Rehborn 1998). The FOTO and ASDA models recognize and track the synchronized flow and wide moving jam traffic phases in congested traffic based on traffic measurements.

Within the FOTO model, traffic phase identification is performed first. Secondly, FOTO performs the recognition of the locations of the upstream and downstream fronts $x_{up}^{(syn)}$, $x_{down}^{(syn)}$ of synchronized flow. While the downstream front $x_{down}^{(syn)}$ of synchronized flow is fixed at the bottleneck, the upstream front $x_{up}^{(syn)}$ is calculated by a “cumulative flow” approach with the inflowing vehicles $q_0(t)$ (vehicles/h) and the vehicles escaping downstream from the synchronized flow $q_n(t)$ (vehicles/h)

$$x_{up}^{(syn)}(t) = \mu \frac{1}{n} \int_{t_0}^t (q_n(t) - q_0(t)) dt, \quad t \geq t_0 \quad (3)$$

with $20 < \mu < 40$ (m/vehicles) as parameter and n as the number of freeway lanes.

Then, the ASDA model recognizes the upstream and downstream fronts $x_{up}^{(jam)}$, $x_{down}^{(jam)}$ of wide moving jams. The positions of these fronts are calculated by

$$\begin{aligned} x_{up}^{(jam)}(t) &= \int_{t_0}^t v_{gl}(t) dt \\ &\approx - \int_{t_0}^t \frac{q_0(t) - q_{min}}{\rho_{max} - (q_0(t)/v_0(t))} dt, \quad t \geq t_0 \end{aligned} \quad (4)$$

$$\begin{aligned} x_{down}^{(jam)}(t) &= \int_{t_1}^t v_{gr}(t) dt \\ &\approx - \int_{t_1}^t \frac{q_n(t) - q_{min}}{\rho_{max} - (q_n(t)/v_n(t))} dt, \quad t \geq t_1 \end{aligned} \quad (5)$$

with $q_0(t)$, $q_n(t)$ as the upstream (downstream) flow rates, $v_0(t)$, $v_n(t)$ as the upstream (downstream) vehicle speeds, q_{min} as the flow rate in the wide moving jam (normally set to zero), and the parameter ρ_{max} as the density inside the wide moving jam. The densities ρ_{max} and $\rho_{min} = q_n(t)/v_n(t)$ in case when free flow is formed downstream of the jam are the two distinct characteristic densities that determine the velocity of the downstream front of a wide moving jam (Kerner 2008).

The formulas (4) and (5) are associated with the classic Stokes shockwave formula as follows:

$$v_s = \frac{q_2 - q_1}{\rho_2 - \rho_1} \quad (6)$$

with q_1 , ρ_1 as the related flow rate and density downstream of the shockwave and q_2 , ρ_2 as the related flow rate and density upstream of the shockwave. It must be stressed that this Stokes formula (6) and the related ASDA formulas (4) and (5) are fundamentally different from the well-known formula for shockwaves in the classic Lighthill-Whitham-Richards (LWR) theory (Whitham 1974):

$$v_s^{(LWR)} = \frac{Q(\rho_2) - Q(\rho_1)}{\rho_2 - \rho_1} \quad (7)$$

where $Q(\rho_2)$ and $Q(\rho_1)$ are flow rates on the fundamental diagram that gives a single correspondence between the density and the flow rate. In contrast, in Eq. 6 and ASDA formulas (4) and (5), there is no given correspondence between the density ρ_2 and the flow rate q_2 . ASDA solves the problem to find q_2 through the use of measured data from an upstream detector. For each time interval (e.g., 1 min) based on measured data, ASDA formulas (4) and (5) find the density as a ratio of the flow rate and the speed. As shown in empirical observations in synchronized flow, there is no single relationship between flow rate and density. Instead there are an infinite number of such relations covering a two-dimensional region in the flow-density plane (“2D states” are associated with Kerner’s hypothesis). In other words, in contrast to the LWR theory, there is no fundamental diagram of congested traffic that gives a single flow rate for a given density or vice versa a single density for a given flow rate. For this reason, as we mentioned in the hypotheses of Kerner’s three-phase traffic theory, there is a 2D region of steady states in synchronized flow.

The front locations $x_{up}^{(syn)}$, $x_{down}^{(syn)}$ and $x_{up}^{(jam)}$, $x_{down}^{(jam)}$ define the spatial size and location of the related “synchronized flow” and “wide moving jam” objects, respectively. Finally, the FOTO and ASDA models track these object fronts $x_{up}^{(jam)}(t)$, $x_{down}^{(jam)}(t)$, $x_{up}^{(syn)}(t)$, $x_{down}^{(syn)}(t)$ in time and space (see Fig. 29). Note that through the use of the FOTO and ASDA models, the tracking of congested traffic objects is also carried out between detectors, i.e., when the object fronts cannot be measured at all. Additionally, the FOTO and ASDA models perform without any validation of model parameters in different environmental and traffic conditions. Model applications are not limited to stationary detector measurements which could measure the necessary flow rates and vehicle speed directly; the use of more advanced measurement technologies like floating car data (vehicles acting as moving traffic sensors) or phone probes (phones acting as moving traffic sensors) is also possible.

Tracking of spatiotemporal congested patterns at freeway bottlenecks is done for 1200 km of freeway network in the state Hessen (Germany) through the use of the FOTO and ASDA models in an online application in the traffic control center and evaluated comprehensively in a 3-month investigation of all measured congestion on the freeway A5 (Kniss 2000). Since April 2004, measured data of nearly 2500 detectors are automatically analyzed by FOTO and ASDA models (Kerner 2004; Kerner et al. 2004). The resulting spatiotemporal traffic patterns are illustrated in a space-time diagram showing congested pattern features.

Figure 30 shows the space-time diagram for very large traffic congestion at the A5-North in Hessen reconstructed from detector data by FOTO and ASDA models. On the 14 June 2006, for about 18 km and nearly 7 h, an “expanded pattern” exists on the freeway. The numbers 1–4 illustrated in Fig. 4 represent the trajectory for individual drivers at different times of the measurement period with different travel times for each driver on the 20 km section: the first one starts at 13:30 h in free traffic and passes in 10 min, the second starts 1 h later and needs approx. 80 min (in heavy congested traffic), the third starts at 17:30 h and it takes approx. 85 min to pass this section, and the fourth driver starts at 19:30 h and needs approx. 65 min to drive through the congested pattern in the decreasing congestion. The delay times of more than 1 h are caused by the reduced vehicle speeds in both the “synchronized flow” and “wide moving jam” traffic phases, respectively.

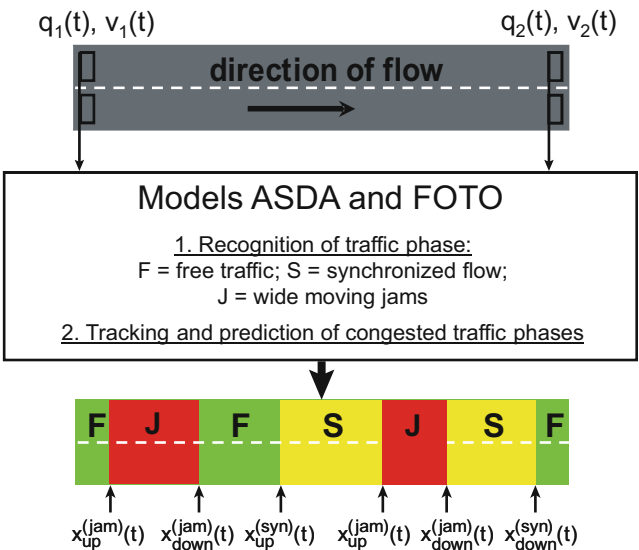
The traffic control center in Hessen with its model application of FOTO and ASDA models is the laboratory environment for first realization of freeway traffic prediction based on three-phase traffic theory, because of online and archived detector measurements, historical time series, and the recognition of the current situation by FOTO and ASDA models.

Detection of Current Traffic States with Floating Cars

Floating cars can send information on their individual drive based on predefined protocols (e.g.,

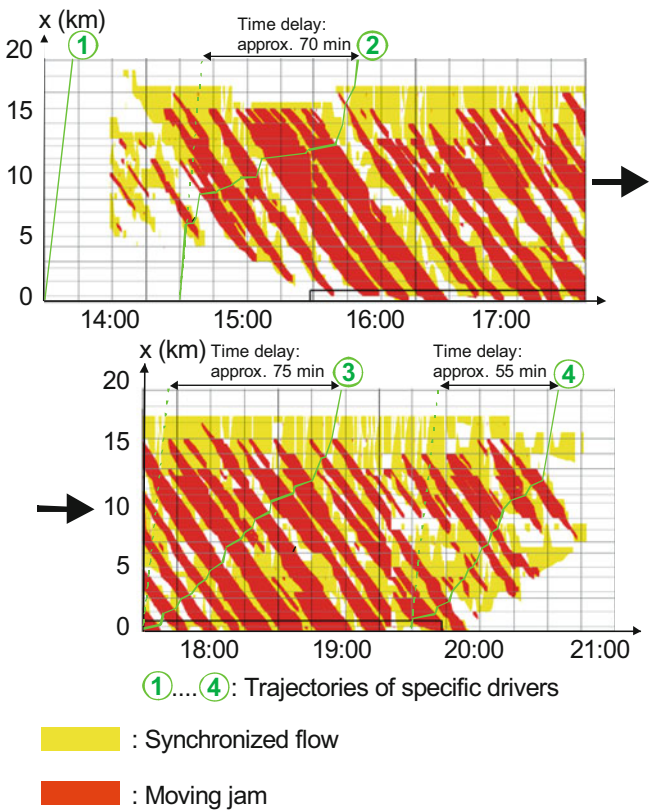
Traffic Prediction of Congested Patterns,

Fig. 29 Illustration of the FOTO and ASDA model approach (Kerner 1999b, 2004, 2008; Kerner and Rehborn 1998; Kerner et al. 1998, 2004)



Traffic Prediction of Congested Patterns,

Fig. 30 Congested traffic pattern reconstructed by FOTO and ASDA models: space-time diagram with vehicle trajectories 1–4 and related delay times from freeway A5-North in Hessen, 14 June 2006



GTP: Global Telematics Protocol, www.telematicsforum.com). A specific onboard algorithm interprets the vehicle data and decides if an event criterion like a slow speed for a certain time interval of a freeway is recognized (Becker and Fastenrath 1998; Fastenrath 1998). Then, a part of the vehicle trajectory is communicated to the traffic control center. The vehicle can act as a moving sensor in traffic in addition or even as a replacement to stationary roadside sensors of the infrastructure (e.g., Boker and Lunze 2001).

The empirical Figs. 31, 32, and 33 show the vehicle data solely and in combination with FOTO and ASDA results from the traffic control center based on stationary detectors (horizontal lines in Figs. 31, 32, and 33) (see Rehborn et al. 2003, for further examples). Several vehicles have sent their messages on this 20 km freeway stretch. The colors of the vehicle trajectories illustrate the different speeds. Solely based on this vehicle information, the blue dotted locations of congested patterns, i.e., a region of synchronized flow and two propagating wide moving jams, would have been assumed and registered in the traffic control center. In Fig. 31 the existence of

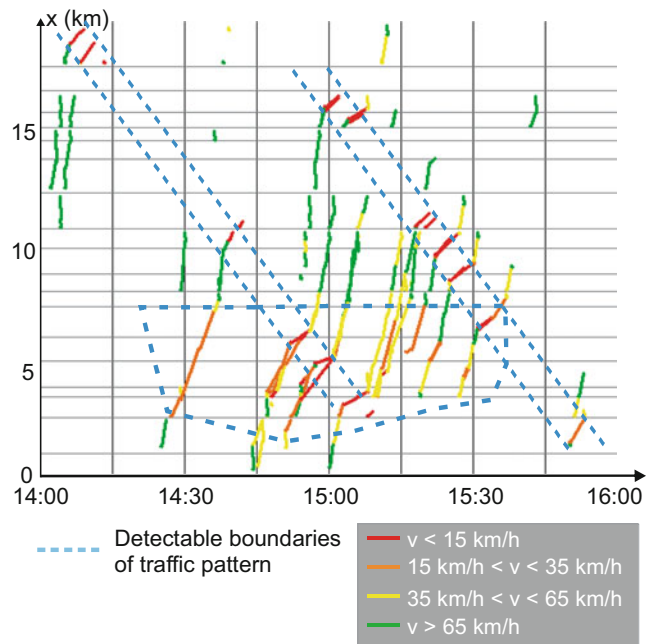
spatiotemporal traffic patterns in congested traffic has been detected solely based on moving detectors, i.e., moving vehicles. The vehicles detect in this congested situation one traffic phase with a fixed downstream location (blue region between 2 and 7 km in Fig. 31) and two examples of a propagating traffic phase in congested traffic (blue bars diagonal on the whole freeway stretch in Fig. 31). As a consequence, the vehicles have measured empirically two different traffic phases in congested traffic as it is stated in Kerner's three-phase traffic theory (Kerner 2004).

The number of vehicles, which have sent messages in this example, has been about 20 per hour which correlates to a penetration rate of about 0.5% (estimated 4000 vehicles/h on the two lane freeway). These quantitative results give the impression that such a small penetration rate of equipped vehicles could be sufficient for congested traffic state recognition on freeways without any stationary sensors.

The same traffic situation has been reconstructed by FOTO and ASDA models solely based on local detectors (Fig. 32). Qualitatively the same regions of the two propagating wide moving jams and the

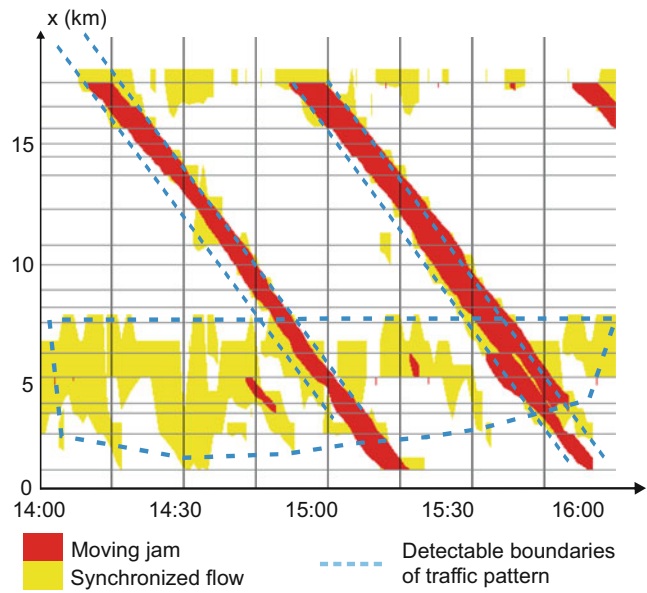
Traffic Prediction of Congested Patterns,

Fig. 31 Vehicle trajectories (approx. 20 per hour) from GTP equipped vehicle (event messages per vehicle have been sent to the center) (See Rehborn et al. 2003)



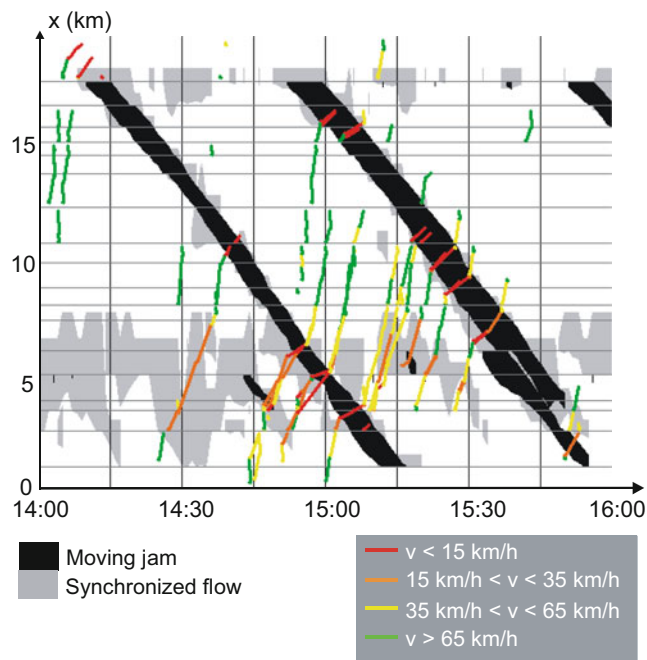
Traffic Prediction of Congested Patterns,

Fig. 32 Congested pattern reconstruction with FOTO and ASDA models (See Rehborn et al. 2003)



Traffic Prediction of Congested Patterns,

Fig. 33 Overview of FOTO and ASDA traffic reconstruction and vehicle trajectories (See Rehborn et al. 2003)



synchronized flow have been recognized (i.e., the blue dotted regions).

The overview of all available data sources, i.e., local detectors with FOTO and ASDA reconstruction and vehicle trajectories, proves the technical feasibility of detecting congested traffic patterns with different measurement techniques, both local

and moving like a vehicle. It has to be taken into account that based on only one vehicle it is impossible to reconstruct the whole picture of the congested traffic. To perform such a congested traffic pattern reconstruction inside a vehicle, the vehicle has to have additional information from other drivers, beacons, or traffic centers.

Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner's Three-Phase Traffic Theory Versus Other Approaches

As explained in Introduction, from a methodological point of view, two main traffic prediction concepts can be distinguished (see Fig. 2):

- (i) "Data mining": statistical methods based on regression (e.g., Davis and Nihan 1991; Sun et al. 2003; Williams 2001; Williams and Hoel 2003; Yang et al. 2004b), wavelet (e.g., Jiang and Adeli 2004; Sun et al. 2004), or filtering models as well as neural networks (or in hybrid combinations) (e.g., D'Angelo et al. 1999; Pancratz 1991)
- (ii) "Physics of traffic": approaches with the use of empirical (measured) spatiotemporal features of traffic congestion as well as traffic flow modeling (e.g., Acha-Daza and Hall 1993; Ben-Akiva et al. 2003; Chrobok et al. 2001; Daganzo 1994, 1995, 1999; Nagel and Schreckenberg 1992; "Traffic Flow Theory 2006" 2006; Wang and Papageorgiou 2005; Wahle et al. 2000.

We have also stated in Introduction that the first kind of approaches more or less take techniques from other disciplines like mathematics, information technology, or artificial intelligence and try to transfer successful concepts into traffic science. The nature of empirical traffic process does not play a key role in the data mining approaches. The "data mining" techniques have the principal problem of determining which data features they should focus on and which are less or not relevant for the accurate prediction.

The "physics of traffic" approaches are based on the nature of traffic phenomena. They begin with the understanding of measurements of traffic variables made in space and time as well as with simulations of driver's behavior, roads with possible bottlenecks, etc. The key element for a successful prediction concept here lies in the correct and reliable understanding of the empirical features of the traffic process and their reproducible characteristics. Furthermore, this understanding

of real traffic should be incorporated in a traffic flow model, which should explain and predict the empirical traffic features.

It is well known that traffic flow is subject to occasional abrupt disturbances (e.g., incidents, congestion) that can potentially change the underlying dynamics and the stability of the data (Disbro and Frame 1989). For example, when a traffic accident occurs, the average speed may suddenly drop to a new level. Not only discrete events like accidents can alter the underlying dynamics but also the level of congestion. Nevertheless, reliable predictions are especially valuable and needed on occasions when unexpected events occur (Lieu 2000). In fact, researchers also acknowledge the occurrence of larger prediction errors in their models whenever traffic congestion sets in and dissolves (Van Lint et al. 2002; Van Lint and Van der Zijpp 2003) and when accidents occur (Zhang and Rice 2003).

Many researchers have investigated the field of traffic science (see Cremer 1979; Daganzo 1997; Helbing 1997; "Highway Capacity Manual 2000" 2000; Kerner 2004; Leutzbach 1988; May 1990; Papageorgiou 1983; Schönhof and Helbing 2007; "Traffic Flow Theory 2006" 2006). However, due to a lack of understanding of the underlying physical processes and the availability of the relevant high-quality traffic data as well as its interpretation, applications of traffic prediction are today still in their infancy. In Kerner's three-phase traffic theory (Kerner 2004), empirical phenomena on freeways have been explained well enough to develop reliable methods for traffic prediction. In this paper, traffic prediction methods are based on recurring and predictable features of traffic congestion as they are empirically proven in the three-phase traffic theory. The combination of empirical features of freeway traffic and its correct interpretation with a pattern database is the key for successful and precise traffic prediction.

Successful traffic prediction should be focused more or less on a congestion prediction. In the industrialized world with heavy vehicle traffic on roads, the congestion is still of much less duration than the free traffic: a very simple traffic prediction citing "free traffic" at all times on all roads would be correct in large parts of the networks.

The time basis for traffic prediction evaluation must not cover the whole day, but the duration of the relatively seldom congestion events. The traffic prediction quality should be evaluated and focused on for the congested time periods of the day. Prediction errors of about 30% are reported for those heavily congested situations (Ishak and Al-Deek 2002). The deviations of the “real” congested situation on the road and the predicted situation have to be compared later on in the laboratory to evaluate the prediction approach. It has to be taken into account, that in today’s measurement systems, the detection of the reality in freeway networks is still imperfect, but those results of a detection system have to be used as a “gold standard” for traffic prediction evaluation.

Definition of “Short-Term” and “Long-Term” Traffic Predictions

Freeway traffic prediction could additionally be differentiated based on the time horizon: the short-term prediction in a horizon of hours can be influenced by the current situation on the road. If larger time horizons have to be forecasted in traffic, the traffic prediction should be based more on historical assumptions and less on the current situation.

In literature, there is no single definition which allows distinguishing between short-term and long-term predictions. Sometimes, traffic prediction on a current day is considered as a short term and traffic prediction on the next days as long term. In other publications, short term is a prediction during a chosen time interval, e.g., that is equal or shorter than 60 min (Smith and Oswald 2003). Overviews on short-term prediction approaches can be found in, e.g., Arem et al. (1997), Lee et al. (1998), Lindveld et al. (2000), Middelham (2001), Smith and Demetsky (1997), Smith et al. (2002), Vlahogianni et al. (2004), Zhang et al. (1998), and Zhang and Rice (2003). Short- and long-term prediction can also be distinguished through the use of current measured data: if in the traffic prediction the current measured data are taken into account, then the prediction is considered as short term, because this prediction has a sense for the current day only. Traffic prediction in which only historical time

series are used is then long-term prediction. The prediction for the next days is automatically a long-term prediction, because current measured data have usually no influence on prediction for the next days. In general distinguishing between long- and short-term predictions made in the article is used for clarity of consideration rather than for analysis of the related prediction methods. Indeed, many of the methods discussed below can be used both for short-time and longtime predictions.

A possible definition of “short term” could be based on the physical features of the congested patterns found in Kerner’s three-phase traffic theory and the methodology of the prediction approach. If the prediction of the congested pattern is influenced by the current situation on the freeway, i.e., there is already a congested traffic pattern emerging, this situation is called “short-term prediction.” If there is no use of the current traffic situation in the traffic pattern prediction method, the traffic prediction will be called “long-term prediction,” because it is performed on any kind of historical assumption of breakdown probabilities or traffic pattern correlations and not the current situation. With this definition, a short-term prediction in contradiction to a solely time-based definition as in Smith and Oswald (2003) could last several hours (e.g., for more than 5 h in Fig. 30 after the pattern emergence).

Traffic Flow and Travel Time Prediction

Many of the proposed forecasting models are focused on the traffic flow (e.g., Abdulhai et al. 1999; Chen and Grant-Muller 2001; Chen and Chien 2001; Ding et al. 2002; Hamed et al. 1995; Head 1995; Jiang and Adeli 2004; Lan and Miaou 1999; Lu 1990; Nicholson and Swann 1974; Okutani and Stephanedes 1984; Pinkofsky 2002; Smith and Demetsky 1994, 1997; Smith et al. 2002; Smith and Oswald 2003; Stathopoulos and Karlaftis 2003; Wild 1997; Williams 2001; Williams and Hoel 2003). Furthermore, most of these approaches predict the traffic flow conditions at a location given previous observations at that point only. This neglects the spatiotemporal behavior of traffic with upstream moving structures. As proposed by Williams

(2001), an efficient forecasting approach should involve the use of traffic flow data from upstream detectors to improve the forecasting at downstream locations. The free traffic is flowing downstream, but the congestion is propagating upstream, as will be explained later in more detail; therefore the flow rate upstream is influenced by the congestion emergence downstream which is very relevant for successful predictions. Although travel times have been identified as an important parameter for the real-time performance of transportation management systems, most elder studies have been focused on predicting traffic flows. This is mainly because most traffic surveillance systems, especially in the USA, use single inductive loop detectors. These have point-measured information regarding traffic flow and occupancy. For many applications, the travel time is the most useful prediction variable, and therefore much research effort has been dedicated to predicting travel times (Chen and Chien 2001; Chien and Kuchipudi 2003; D'Angelo et al. 1999; de Rham and Lange 2000; Dharia and Adeli 2003; Fallah-Tafti 2001; Huysken and Van Berkum 2003; Kaysi et al. 1993; Kisgyorgy and Rilett 2002; Kuchipudi and Chien 2003; Kwon et al. 2000; Lee et al. 1998; Lindveld et al. 2000; Matsui and Fujita 1998; Nam and Drew 1996; Nanthawichit et al. 2003; Nicholson and Swann 1974; Nikovski et al. 2005; Oda 1990; Ohba et al. 1997; Park et al. 1999; Petty et al. 1998; Rice and Van Zwet 2001; Rilett and Park 2001; Schrader et al. 2004; Van Lint et al. 2002; Van Lint and Van der Zijpp 2003; Zwahlen and Russ 2002). Those travel times are very relevant as static input data for trip planning, i.e., the reliable dynamic adaptation of navigation systems (Chen et al. 2007).

Statistical Approaches and Time Series

Other known methods (Nikovski et al. 2005) use approaches from mathematical statistics to “learn” coefficients of traffic prediction in the form of correlations in time and space. These ideas try to learn implicitly spatiotemporal features of the traffic process and try to find correlations between heterogeneous traffic data for time series prediction. For example, the velocity of the propagating wide moving jam of 15 km/h against the direction

of the flow as characteristic parameter of this traffic phase (Kerner 2004) can be “learned” statistically: such a wide moving jam registered downstream could be predicted with high probability to be 3 km upstream in about 12 min.

Another widely used approach is the use of time series of traffic variables (e.g., Ding et al. 2002; Pinkofsky 2002; Van der Voort et al. 1996; Vlahogianni et al. 2006). Average flow rates and speeds vary over time but are regular and recurring for the same network location. Therefore, clustering methods based on time series similarities are used for traffic prediction. Such average flow rates can be useful for an estimation of the expected congestion due to tomorrow's roadwork on a freeway section. Here the average flow rate is taken as traffic demand estimation. One of the principal problems of these clustering methods is the fact that the congested situations are part of the measurements which are used for the traffic time series. Therefore, if congestion is part of a flow rate time series at a specific location or section, the demand is underestimated because no vehicles have passed the detector but have had to wait in the congested area further upstream: for prediction the expected flow rate is systematically underestimated.

Several authors have used statistical approaches (e.g., Chickering et al. 1997; Horvitz et al. 2005; Moorthy and Ratcliffe 1998; Nicholson and Swann 1974; Nihan and Holmesland 1980; Nikovski et al. 2005; Pancratz 1991; Pinkofsky 2002; Schrader et al. 2004; Williams and Hoel 2003) and time series for traffic prediction. The spikes of travel time curves caused by (sudden) traffic breakdowns are a problem for these kinds of linear approaches, because they are underrepresented in the data set and the regression filters and smooths the traffic breakdowns. Finding a proper prediction model defining the prediction variable and the number of necessary data cycles is a question of finding adequate parameters. Therefore, the success of regression techniques for short-term congestion predictions seems questionable.

Neural Networks

One widely used “data-driven” approach was the use of artificial neural networks (ANNs) for traffic

prediction. As a relatively new mathematical model, ANNs were first introduced in the 1940s (e.g., Haykin 1999; Hecht-Nielsen 1990; Rumelhart and McClelland 1986). The basic idea of the neural network is to emulate the human brain, which consists of a very large number of interconnected neurons. Each of the neurons (“cells”) have similar behavior, and there is the possibility of having several layers of neurons in an ANN. ANNs try to adopt features of a given traffic data set in a learning phase and then perform a traffic prediction based on a given data set (i.e., the current traffic situation). Many researchers have used ANN for data completion or short-term forecasting of traffic variables which very often focus on travel times (Abdulhai et al. 1999; Chen et al. 2001; Dharia and Adeli 2003; Dia 2001; Dougherty 1995; Fallah-Tafti 2001; Huang and Ran 2003; Innama 2001; Ishak and Alecsandru 2004; Kirby et al. 1997; Kisgyorgy and Rilett 2002; Lingras et al. 2002; Matsui and Fujita 1998; Park et al. 1998, 1999; Rilett and Park 2001; Smith and Demetsky 1994; Van der Voort et al. 1996; Van Lint et al. 2002; Xiao et al. 2003; Yang et al. 2004a; Yasdi 1999; Yin et al. 2002; Zhang et al. 1998; Zhang 2000). ANNs try to learn and predict spatiotemporal relationships in traffic data. They are capable of dealing with the spatiotemporal relationships implicitly without understanding the underlying complex processes. The ANN predictions are a “black-box” process and, however, depend strongly on the completeness of the input data set. Therefore, the success of the ANN approach rises and falls with the choice of the training data set, which makes the approach difficult in prediction realization.

Pattern Matching: Fitting Traffic Pattern from Database

As it is said before, the congested traffic patterns SP, GP, and EP with its variants on freeways are composed of two basic elements (“traffic phases”), i.e., the regions of wide moving jams and of synchronized flow. The prediction will be done by forecasting the current spatiotemporal traffic objects, i.e., the regions of the wide moving

jams and synchronized flow (Kerner 2004). Position and width have to be predicted: this is performed based on the most similar traffic pattern for the related effectual bottleneck in the pattern database. The traffic prediction is calculated as follows: firstly, the current situation is reconstructed by FOTO and ASDA models. Next, a filtered current (“generalized”) traffic pattern is used as a search pattern for the traffic pattern database. The clustering is carried out to generalize the traffic pattern into some basic “puzzle” pieces of wide moving jam and synchronized flow objects. The reason for this is the dependence of the traffic pattern reconstruction quality on the detector system, which often has limitations in precision and reliability. Therefore, a congested traffic pattern measured by inductive loop detectors and interpreted by FOTO and ASDA models like in Fig. 30 is clustered and combined into larger regions of synchronized flow and wide moving jams, respectively. The similarity between the current clustered pattern and the pattern database is determined based on a distance measure which allows deviations in space and time (e.g., some hundreds of meters and some minutes as parameters). Then, the traffic pattern from the database is linked to the current pattern. The aggregated predicted delay time of the congested pattern consists of the delay time contributions of (i) the propagation of the current wide moving jams already detected, (ii) the predicted regions of the synchronized flow from the traffic pattern database, and (iii) the predicted new wide moving jams emerging in the future as information from the traffic pattern database.

If traffic congestion has been detected, how can the prediction pattern be identified from the database? Criteria for this “matching” process, i.e., finding the most probable traffic pattern for the future, are freeway, location of the effectual bottleneck, time, size of the current synchronized flow region, location of the first wide moving jam (if GP or EP), and the number and frequency of the wide moving jams (if GP or EP).

For applications, it should be noted that the matching process could use the current traffic situation from a traffic control center (e.g., based on online traffic reconstruction with FOTO and

ASDA models) or from autonomous vehicles based on the vehicle internal measurements (Fig. 34).

Prediction of Propagation of the Fronts of Wide Moving Jam

An approach for wide moving jam short-term prediction is as follows: from three-phase traffic theory (Kerner 2004), the upstream stable propagation of the downstream front of a wide moving jam is a characteristic parameter of traffic. The upstream front is influenced by the incoming traffic flow into the wide moving jam. A short-term prediction of a wide moving jam can be based on this stable propagation feature: if a wide moving jam is detected and registered by ASDA model, the mean velocities of the fronts can be used for short-term predictions of the wide moving jam. After calculation of the mean velocities of the fronts, those values are used to predict the wide moving jam's position in the future. Therefore, for the wide moving jams, stable movement in space and time can be calculated based on the registration times at the local detectors.

A short-term prediction of wide moving jams is suggested in Kerner et al. (1998). If both fronts of a wide moving jam have been registered, the clearing up of the jam can be obtained. This could be useful to predict the time, when traffic control methods regarding the wide moving jam could be eliminated. The time t_{clear} is obtained as (Kerner et al. 1998):

$$t_{\text{clear}} = t_{\text{reg}} - \frac{\int_{t_0}^{t_{\text{reg}}} v_{\text{gl}}(t) dt}{v_{\text{gl}}(t_{\text{reg}}) - v_{\text{gr}}(t_{\text{reg}})} \quad \text{at} \quad |v_{\text{gl}}| < |v_{\text{gr}}| \quad (8)$$

with v_{gl} and v_{gr} as the related velocities of the up- and downstream front of the wide moving jam, respectively. t_0 , t_{reg} are the registration times of the upstream and downstream front of the wide moving jam at the downstream position. Figure 35 illustrates scheme and empirical example for a wide moving jam clearing up.

A short-time travel time prediction can be established in the case of a propagating wide moving jam, if the downstream front of the wide

moving jam has been registered at a downstream location. As illustrated in Fig. 36, during the first drive section until the upstream front reached at t_{up} , the drive line is based on an average free speed v_0 of the vehicle. For the subsequent drive section within the wide moving jam, the average vehicle speed of v_{jam} is used to generate the drive line. The last drive section after the downstream front of the wide moving jam is then calculated with the speed v_n . Therefore, the individual vehicle travel time for this section (see Fig. 36) can be obtained as (Kerner et al. 1998)

$$t_{\text{R}} = \frac{L + v_n t_{\text{down}} - v_0 t_{\text{up}} - v_{\text{jam}}(t_{\text{down}} - t_{\text{up}})}{v_n} \quad \text{at} \quad t_0 = 0 \quad (9)$$

In many cases, the remaining speed within the wide moving jam v_{jam} can be neglected, so that for the travel time the following approximation formula is obtained (Kerner et al. 1998):

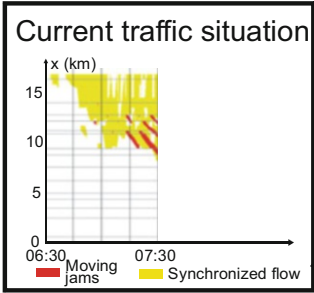
$$t_{\text{R}} = \frac{L + v_n t_{\text{down}} - v_0 t_{\text{up}}}{v_n} \quad \text{at} \quad t_0 = 0 \quad (10)$$

This method can also be used when several wide moving jams occur between two measuring points. In this case, assuming that no entry or exit roads exist in the monitoring area, the plausible assumption used is that the flow and average speed of vehicles corresponds both upstream and downstream of the wide moving jam at the time after the downstream front has passed the downstream measuring point (Kerner et al. 1998).

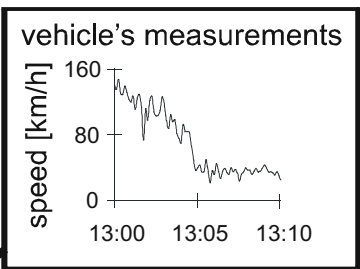
Traffic Pattern Prediction with Traffic Flow Models

Traffic flow models offer a possibility to perform traffic predictions in different scenarios (e.g., Chrobok et al. 2001; Daganzo 1994, 1995, 1999; Maerivoet and De Moor 2005; Nagel and Schreckenberg 1992; Wahle et al. 2000): through the use of a model, different (possibly future) influences on traffic, e.g., incidents, roadwork, accidents, etc., could be simulated. Therefore, the consequences of such future events could be

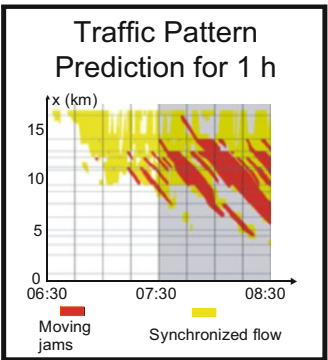
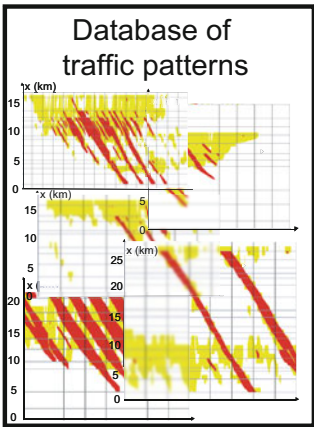
Traffic state reconstruction
e.g., by ASDA/FOTO
based on detectors



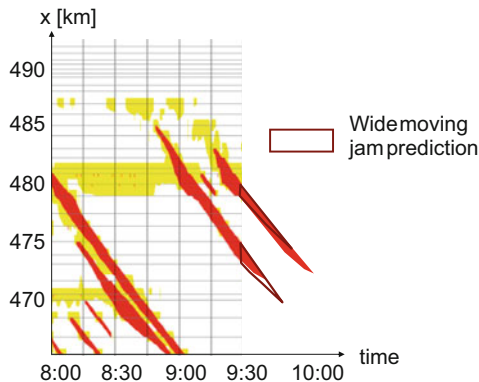
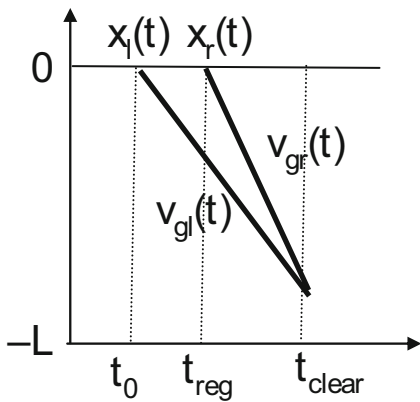
Traffic state reconstruction
by vehicle measurements



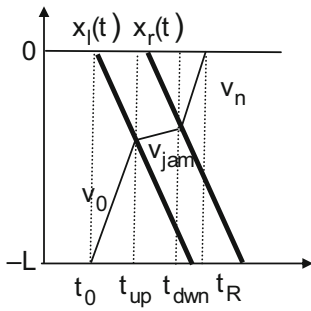
**Pattern-Choice:
"Matching"**



Traffic Prediction of Congested Patterns, Fig. 34 Use of spatiotemporal traffic pattern database for traffic prediction with “matching” (Kerner 2004)



Traffic Prediction of Congested Patterns, Fig. 35 Scheme for a wide moving jam clearing up (left, scheme; right, empirical example) (Kerner et al. 1998)



Traffic Prediction of Congested Patterns, Fig. 36 Scheme for a travel time prediction for a vehicle (Kerner et al. 1998)

estimated and used for current decisions, e.g., the choice of the best traffic management strategy. In this context, a traffic prediction based on traffic modeling could be a valuable instrument.

Many authors have used traffic flow models for traffic prediction, e.g., Chrobok et al. (2001), Daganzo (1994, 1995, 1999), Nagel and Schreckenberg (1992), and Wahle et al. (2000). One of the key requirements for successful modeling is the correspondence with the empirical features of freeway congestion. In Kerner and Klenov (2002, 2003, 2006) it is shown that all earlier traffic flow model approaches like the abovementioned ones based on the so-called “fundamental-diagram” approach are not able to explain those empirical features. As a consequence, the related traffic prediction results are doubtful.

On the other hand, traffic prediction and scenarios modeled based on KKW model (Kerner 2004) could be a reliable instrument for decision support with traffic simulation: the consequences of different influences on traffic flow could be simulated in advance. More detail information on traffic simulation can be found in the paper (Kerner 2008) of this encyclopedia.

Long-Term Traffic Prediction

Many empirical studies of traffic have shown that the time dependence of traffic flow possesses some predictable features (e.g., Heidemann and Wimber 1982; “Highway Capacity Manual 2000”

2000; Kerner 2004; Kerner et al. 2005; May 1990; Ober-Sundermeier and Zackor 2001; Pinkofsky 2002) because they are recurring and regular. These features depend on characteristics of the day (e.g., working day or weekend) or whether there is an event like a soccer match. The regularity and recurrence are even stronger if a certain time shift of the congestion emergence is allowed, the morning traffic peak with the same traffic congestion 10 min later is not an “anomaly” but still a regular and similar traffic pattern.

If the characteristics of a day are known, there is a certain set of dependences of the flow rates over time for a specific freeway section. Each of these dependences is one of the historical time series for flow rates. A historical archive of traffic flow rates is therefore the basis for a time series database structured by the characteristics of the different days. At least two parameters of the databases are the day characteristics and an event which influences the traffic. The historical database contains a related probability of the occurrence for each of the time series for each of the different parameters. A cluster analysis of the historical data is a common approach to produce the time series database.

For urban areas, time series-based predictions have been proposed (e.g., Hoyer et al. 2003; Kerner 2004; Yin et al. 2002). Those methods should predict the traffic state for different time horizons. The database should use data from the infrastructure (road net, lanes, traffic lights), basic traffic engineering data (e.g., capacities, speed limits), and historical traffic data as expectation values of, for example, the flow rates and current measurements.

Two types of prediction based on long-term time series are possible. Firstly, the choice of the most probable time series for a similar event or day in the future (e.g., soccer event, weekday, or weekend) could be applied. Based on archived traffic data, the database contains time series with additional parameters like day, time of day, event, weather, etc. The prediction for the next day is then performed based on a fulfillment of such criteria. Secondly, the prediction can include “matching of time series,” i.e., the prediction is not necessarily the most probable curve for the

future from the historical database, but based on time series with minimal deviation from the current traffic variables.

Long-term time series of traffic variables have two aspects: (i) processing of the raw data to produce a time series database and (ii) their later usage in a predictive application. For the processing of raw traffic data, approaches from signal processing have been proposed (Grenander 1996; Qiao et al. 2003; Sun et al. 2004; Teng and Qi 2003; Venkatanarayana et al. 2005; Xiao et al. 2003) to extract and detect the relevant events in the data. Others take different and multiple methodologies from statistics (e.g., Box and Jenkins 1976; Brockwell and Davis 2002; Chatfield 2001; Ding et al. 2002; Fuller 1996; Gazis and Knapp 1971) like cluster analysis (Heidemann and Wimber 1982; Pinkofsky 2002), ARIMA (autoregressive integrated moving average) approaches for stationary linear processes (Cetin and Comert 2006; Van der Voort et al. 1996; Williams and Hoel 2003), stochastic methods (Ahmed and Cook 1979; Nihan and Holmesland 1980; Nikovski et al. 2005), or Bayes' Theorem (e.g., Chickering et al. 1997; Horvitz et al. 2005). Additionally, neural network approaches from artificial intelligence have been used for unsupervised learning transforming raw traffic data into time series (Van der Voort et al. 1996; Yasdi 1999).

In general, the methods for long-term time series prediction try to reproduce the traffic phenomena by their different aggregation methods. The resulting time series from the raw data should model a realistic curve of the traffic variable to be useful for a predictive application. The mentioned approaches have a commonality in their usage of a large variety of data analysis techniques. Some of them do not take into account the known physical and predictable features of the traffic process (Kerner 2004), while others even neglect the principle feature of nonlinearity and nonstationarity of traffic (e.g., ARIMA; Van der Voort et al. 1996; Williams and Hoel 2003). Without proper understanding of the relevant features of traffic congestion, it is doubtful that those methods could be optimal and successful.

Also microscopic traffic simulation has been used to estimate long-term effects in traffic in

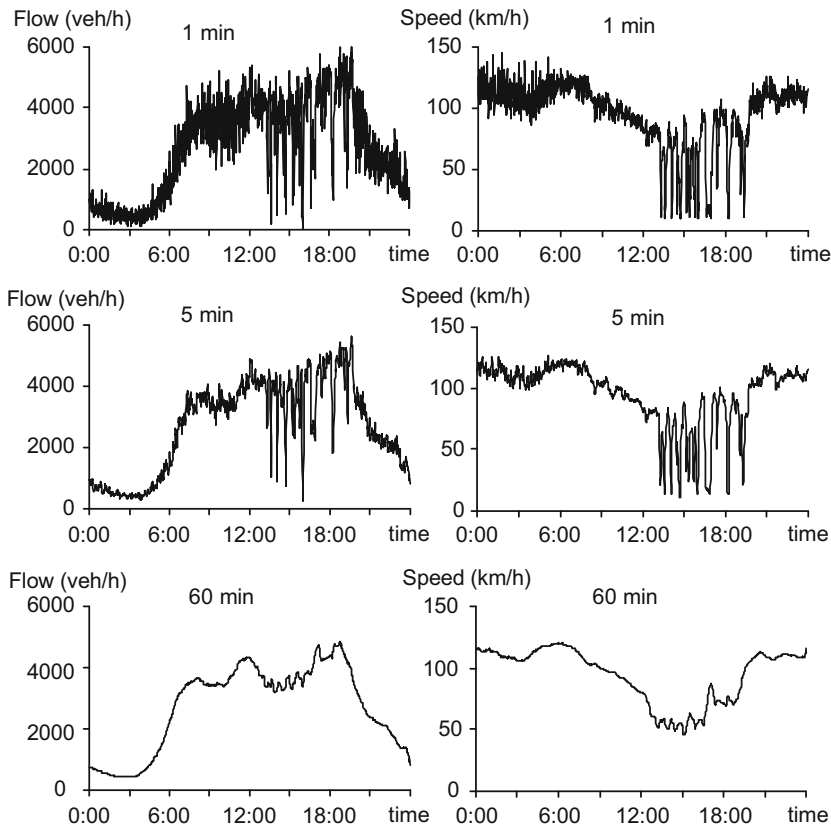
combinations of time series and modeling of traffic process (Chrobok et al. 2001; Kaumann et al. 2000; Kitamura and Kuwahara 2005; Nagel and Schreckenberg 1992; "Traffic Flow Theory 2006" 2006; Wahle et al. 2000). The goal is a network wide traffic prediction in a combination of measurements and modeling techniques.

Predictable Features of Traffic Time Series

Long-term time series of flow rates are used for infrastructure planning and modeling of the traffic demand (e.g., Heidemann and Wimber 1982; "Highway Capacity Manual 2000" 2000; Pinkofsky 2002). For the long term, the typical minute values of traffic data measurements are aggregated in different time intervals. For usage as traffic demand time series, the influence of traffic congestion has to be minimized or filtered. Data aggregation is one key technique to extract predictable features like the flow rate curve from raw measurement data (Oh et al. 2005).

An example of typical detector measurements for double induction loop detectors is shown in Fig. 37. Three loops count the number of vehicles ("flow rate") and their average speed in 1 min intervals on each lane. Here, the aggregated values for the freeway direction A5-North are illustrated for Friday, 23 March 2001. The minute values show several local minima of flow rates in the afternoon, both clearly visible in the flow and speed time series (Fig. 37, top). The emergence of the congestion has occurred at a downstream location, and the local detector on the freeway has registered the local minima which may be associated with a part of the whole congested traffic pattern, e.g., some wide moving jams propagating upstream as part of a general pattern (see Kerner 2004, for explanation): in the dynamic spatiotemporal traffic on freeways, the local detector only registers a part of the whole congested situation). A maximum flow rate of approximately 6000 vehicles/h on this day is a good "rule of thumb" for a three-lane freeway, i.e., 2000 vehicles/h per lane is a good estimation for the maximum possible throughput (including long vehicles).

The moving average of 5 min (Fig. 37, middle) smoothes the fluctuations of the minute interval and shows clearly similar local minima, because



Traffic Prediction of Congested Patterns, Fig. 37 Minute values (*top*), moving averages of this data for 5 min (*middle*) and 60 min (*bottom*) of flow rates (*left*) and

vehicle speed (*right*) from a typical detector from three-lane freeway A5-North between intersections “Nordwestkreuz” and “Bad Homburger Kreuz” on Friday, 23 March 2001

they usually last longer than 5 min. If the moving averaging interval is set to 1 h (Fig. 37, bottom), the resulting time series is smoothed considerably. For long-term applications, e.g., a traffic prediction of the next day, those aggregated hourly traffic data is sufficient. The form of the 60-min time series is typical for a freeway: an increasing flow rate in the morning, higher flow rates during the day, an afternoon peak, and decreasing flows during the evening hours.

Predictable Features of Effectual Bottlenecks

In addition to the temporal aggregation of the previous chapter, a spatial aggregation would enhance the long-term time series prediction to a long-term congested traffic pattern prediction with use of time and location as key predictive parameters. According to three-phase traffic

theory with the defined two phases in congested traffic, i.e., synchronized flow and wide moving jam, spatiotemporal congested traffic patterns emerge after a first phase transition from free flow to synchronized flow that explains traffic breakdown (Kerner 2002, 2004). A bottleneck at which traffic breakdown is observed is called “effectual bottleneck.” It has two empirical features: (i) the phase transition from free to synchronized flow, i.e., traffic breakdown, occurs more frequently at this location in comparison to other freeway locations, and (ii) the downstream front of the synchronized flow region where the vehicles accelerate from the congested region to free flow is almost fixed at this effectual bottleneck (see B_1 – B_3 in Fig. 38). The reason for the existence of an effectual bottleneck is a non-homogeneity of the freeway infrastructure caused by, e.g., on- and

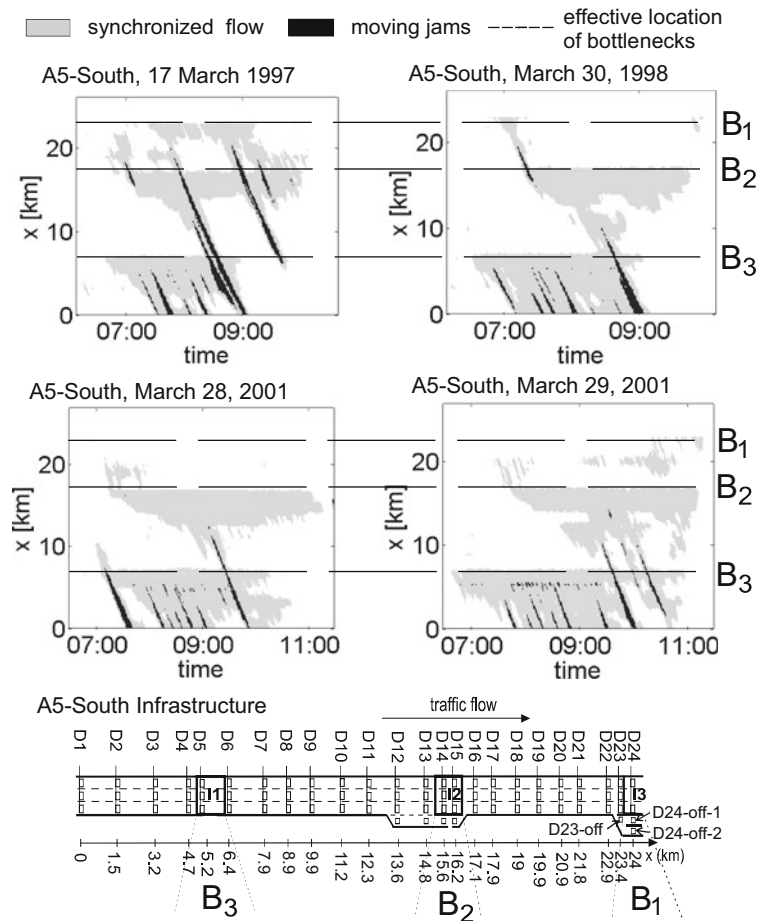
off-ramps, a decrease in the number of lanes, a higher slope, and so on which influences the undisturbed homogeneous traffic flow. From a study of congested pattern formation from 1995 to 2003, it has been shown that the phase transitions from free to synchronized flow at the freeway A5-South have occurred mainly in the vicinity of the effectual bottlenecks B_1 (intersection “Nordwestkreuz”), B_2 (intersection “Bad Homburg”), and B_3 (intersection “Friedberg”) (see Fig. 38).

The resulting congested patterns emerging after the phase transition at the effectual bottlenecks have predictable and characteristic features: the pattern features are the same for different days and years (see Fig. 38 with space over time diagrams). From a global analysis many similarities among the patterns can be identified. The size, form, and duration of the synchronized flow regions, wide moving jam emergence, and their

propagation through different bottlenecks, as well as the time dependences of phase transitions at different bottlenecks (i.e., phase transition from free to synchronized flow always earlier at B_3 than at B_2), are very similar on different days in different years. This supports predictability in a long-term congested traffic pattern database: prediction at a certain time on a certain day could be performed based on the current traffic conditions (current congestion, flow rate, weather, time of day, etc.) and a correlation to the pattern parameters of the database. Furthermore, the data analysis shows that the congested patterns are like “fingerprints” of the road network: each effectual bottleneck may have its specific characteristic parameters of the congested patterns, i.e., type, form, and duration of the congested pattern are linked to the specific bottleneck at which the congestion emerges (Kerner 2004).

Traffic Prediction of Congested Patterns,

Fig. 38 Empirical examples of predictable congested patterns in different years at the effectual bottlenecks B_1 – B_3 on the freeway A5-South (bottom: infrastructure). Overview on regions of free flow (white), synchronized flow (light grey), and wide moving jams (black) in space over time diagrams (Kerner 2004)



For long-term time series of congested traffic patterns, the possible degree of predictability has to be taken into account. According to probabilistic features of the congested patterns, even under the same traffic parameters (e.g., flow rates, weather, road conditions, etc.) the pattern features could be different for different situations (days) with different degrees of predictability. Some deterministic features with a high degree of predictability are as follows (Kerner 2004):

- (i) Location of the effectual bottleneck
- (ii) Emergence of wide moving jams in synchronized flow regions upstream of bottlenecks in a distance of about 3 km or more
- (iii) Stable propagation of wide moving jams through any other traffic states
- (iv) Considerably lower flow rate in the outflow of the wide moving jam if free flow emerges after the jam (about two thirds of the maximum flow rate in free traffic)

Congested pattern features with a middle degree of predictability are:

- (i) The instant of a phase transition from free to synchronized flow at the effectual bottleneck at the same traffic demand (“probabilistic nature of speed breakdown”).
- (ii) The type of congested pattern occurring at an effectual bottleneck. Some of the pattern types (Kerner 2004) may not occur at a specific set of adjacent effectual bottlenecks.

Thirdly, pattern features with a low degree of predictability are:

- (i) The kind of phase transition from free to synchronized flow (induced, i.e., influenced by another adjacent effectual bottleneck or spontaneous, i.e., originally emerging at the related bottleneck)
- (ii) The instant of wide moving jam emergence in synchronized flow

A long-term pattern database therefore should contain the empirical congested pattern features and store the predictable information with related

probabilities. An approach for development of such a congested traffic pattern database is described later on.

Approaches for Producing Long-Term Traffic Time Series

Cluster Analysis for Flow Rates

One kind of detector on freeways and major roads in Germany allows hourly measurements of the flow rates at certain locations in the freeway network (“Dauerzählstellen”). Performing a cluster analysis of this hourly traffic data, typical recurring and regular time series of the flow rates are extracted from the traffic data (Fig. 39). The typical bimodal structure for weekdays (e.g., Monday in Fig. 39) and the unimodal structure for weekend days have been identified clearly (Heidemann and Wimber 1982; “Highway Capacity Manual 2000” 2000; Pinkofsky 2002). In Pinkofsky (2002), seven types of time series for weekdays based on a statistical cluster analysis have been identified which can be characterized by their features:

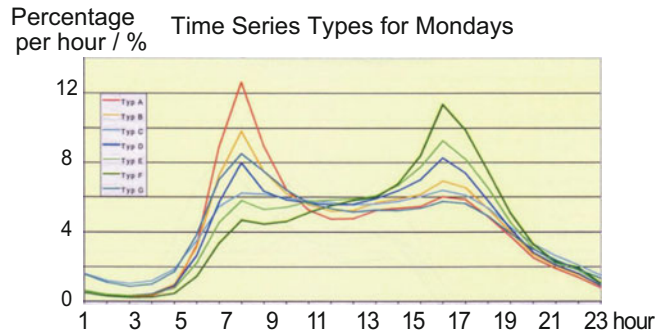
- Type A: strong morning peak
- Type B: morning peak and lower afternoon peak
- Type C: relatively balanced distribution during the daytime
- Type D: double peak
- Type E: afternoon peak, low morning peak
- Type F: strong afternoon peak
- Type G: more than average morning, then decreasing steadily

According to Pinkofsky (2002), Types A and B are dominant on freeways with commuting traffic with the opposite direction of Types E and F.

These time series represent an estimation of the traffic demand for the freeway and major road network; they do not take traffic congestion into account. In a comparison of the averaged detector measurements of Fig. 37 (bottom, left) and the hour values and the clustered results of Fig. 39, the freeway stretch of the A5-North between intersections

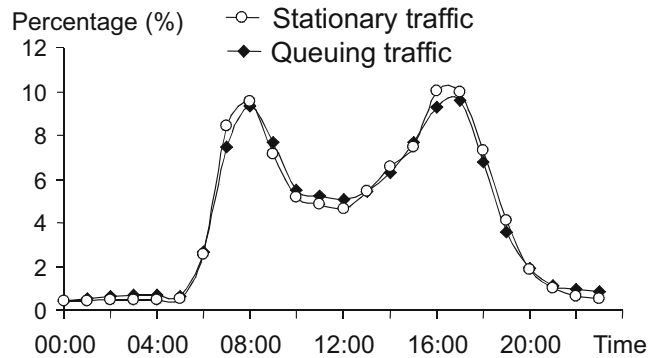
Traffic Prediction of Congested Patterns,

Fig. 39 Daily time series (Weekday Mondays, Type A–Type G) (Pinkofsky 2002)



Traffic Prediction of Congested Patterns,

Fig. 40 Relative percentage of RDS/TMC radio messages (“stationary traffic” and “queuing traffic”) per hour for all days of year 2000 in Germany (Rehborn et al. 2002)



“Nordwestkreuz” and “Bad Homburger Kreuz” on Friday, 23 March 2001 shows a “Type C” form with a relatively balanced distribution during the daytime. In a database, only this type classification and some parameters like the maximum flow rate must be stored for a long-term flow rate time series.

As an additional source of information for long-term time series, the radio traffic messages via RDS/TMC (Radio Data System/Traffic Message Channel) protocol can be used as archived traffic data (Rehborn et al. 2002). The main events of the broadcasted information are the “stationary” and “queuing” traffic which are coded in the broadcast data stream. In the year 2000, approximately one million messages were broadcasted in Germany. A daily distribution of those messages with relative percentages per hour shows in Fig. 40 a double peak in the morning and afternoon hours similar to Type D in Fig. 39. These time series illustrate the time distribution of the traffic congestion; for the whole network, the evening peak is wider than the morning one and

approximately 10–12% of all congestion occur in the freeway network during the two peak hours of the day (from 7:30 to 8:30 h and from 15:30 to 16:30 h, respectively).

The comparison of Figs. 39 and 40 illustrates the correlation between the flow rate time series and the congestion message distribution: from a global perspective, high traffic flows correlate with highly congested traffic states. The interpretation of the RDS/TMC messages of “stationary” and “queuing” traffic is related to speed and flow parameter values as definitions for the TMC protocol, e.g., all measured speeds below 60 km/h are defined as being “queuing” traffic and speeds below 30 km/h are defined as “stationary” traffic. This is in contradiction with empirical features of congested patterns like in Fig. 38, and therefore, those time series cannot be used for successful traffic prediction. The TMC protocol definitions of today make it impossible to have information on the real congested traffic situation, but this information is broadcasted by radio stations, and all radio listeners in most European

(and recently USA) regions are experienced with this common form of traffic information. For dynamic route guidance applications, the “stationary” and “queuing” traffic could be transformed into travel delay times and then be used in navigation systems. A specific high degree of predictability based on the real empirical traffic phases has been lost on this aggregated level of information.

Wavelet Analyses

The extracting and producing of traffic data time series are cumbersome and time-consuming, as well as computer storage intensive. The frequency domain analyses, using Fourier series (Brockwell and Davis 2002), offer a possible alternative, especially regarding storage capacity. The Fourier analysis is appropriate for stationary time series models without spikes. The local minima in traffic variables are on the contrary very important for traffic data analyses because of traffic congestion: they can occur suddenly and are relatively seldom events in comparison to all traffic data measurements. The wavelet analyses try to solve these problems (Burrus et al. 1998; Grenander 1996). Wavelets have been applied to traffic pattern analysis (Jiang and Adeli 2004), traffic prediction (Sun et al. 2004; Xiao et al. 2003), incident detection (Teng and Qi 2003), and determining data aggregation intervals (Qiao et al. 2003) in transportation technology. Wavelet analysis consists of breaking a given signal into other waves of different frequencies, thus forming both a high-pass and low-pass filter at the same time. The ability to analyze the coefficients of the wavelets significantly reduces the dimension of the data vector without losing too much information. This is an efficient approach considering the computational storage costs of the database as well as the performance of the pattern database.

Traffic patterns have some regularity or repetition of some basic underlying structure (Grenander 1996). An example of this regularity is the typical structure of the bimodal daily pattern of weekday traffic with peaks in the morning and in the afternoon (Venkatanarayana et al. 2005) (see Figs. 37 and 39).

For long-term traffic prediction, the traffic data time series are averaged over longer time intervals

(see Fig. 37). Therefore, it seems questionable that the mentioned dominant advantages of wavelet transformation are relevant for long-term time series.

Cluster Analysis for a Database of Congested Traffic Patterns

In an approach supporting Kerner’s three-phase traffic theory concepts (Kerner 2004), a database of congested traffic patterns can be used for long-term predictions. In a first step, the empirical congested pattern has to be clustered. The clustering builds the congested regions into larger spatiotemporal regions and suppresses smaller congested regions. Therefore, a minimum duration of 30 min of the congestion and a width of at least 1 km distance between upstream and downstream front of a congested region are chosen for the clustering approach. Additionally, smaller spatiotemporal gaps have to be filled. Those occur typically in an online detection system: if at a specific location the detector malfunctions or in one cycle the average speed increases a little, there will be one cycle of free traffic within a congested pattern. As a chosen parameter, time gaps up to a 20 min interval will be filled in during the clustering process. The method for both traffic phases is single linkage clustering where the minimum distance from two clusters A and B should be smaller than a parameter X for their merging:

$$\min_{a \in A, b \in B} \{d(a, b)\} < X \quad (11)$$

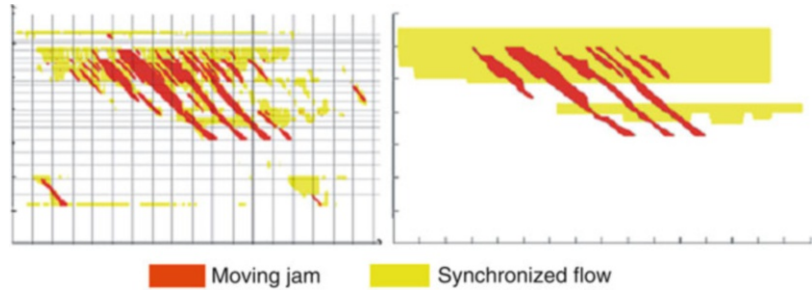
The distance between a and b for two traffic flow cluster is, e.g., calculated by Euclidian distance:

$$d(a, b) = \sqrt{\sum_{k=1}^p (a_k - b_k)^2} \quad (12)$$

The given data points a and b have to be compared for the calculation of the spatiotemporal distance. Figure 41 illustrates the result: from the empirical pattern based on minute values from a detection system and the reconstruction with FOTO and ASDA models (Kerner 2004) on the

Traffic Prediction of Congested Patterns,

Fig. 41 Empirical congested traffic pattern at a bottleneck reconstructed by FOTO and ASDA models (*left*) and clustered traffic pattern for the database (*right*)



left side, the number of wide moving jams is reduced in the clustered traffic pattern on the right. Similarly the regions of the synchronized traffic flow are clustered. The resulting clustered congested pattern could be stored in the database with parameters for both synchronized flow and wide moving jam. In comparison, the clustered pattern can reduce the amount of storage capacity for a congested traffic pattern by about 90% which is very computationally efficient for larger road networks.

In the database of congested traffic patterns, the patterns are stored with some basic parameters which are relevant for each of the traffic phases (“puzzle pieces”). From three-phase traffic theory, it is known that the downstream boundary of the synchronized flow is almost fixed at the bottleneck location. For a synchronized flow region, the starting location and starting time of the synchronized flow region is stored. Secondly, the duration and the expanse are relevant for each of those regions. Two speed parameters describe the form of emergence and dissolution of the synchronized flow regions. The puzzle pieces for synchronized flow have, therefore, four edges in the database.

For the wide moving jams in the traffic pattern database, relevant parameters are the length of the wide moving jams and the maximum time of its propagation life cycle. Additionally, for the frequency of wide moving jams following one after the other, the distance between two following upstream jam fronts is given. To define the angle in the space-time diagram, the speed of the front is stored. The latest wide moving jam of a pattern has to take into account the dissolution of the synchronized flow region, i.e., there is no wide moving jam emerging without a synchronized flow region further downstream. This simplified

database approach is the first step into online traffic prediction application. Furthermore, detailed information of the congested traffic patterns could be stored in prediction applications (Kerner 2004):

- The average speed in synchronized flow region
- The mean velocity and time dependence of the upstream front of the synchronized flow
- Transformations between different types of pattern (SP, GP, EP) after pattern emergence
- Additional information about EP characteristics depending on bottleneck locations
- Probabilities of pattern occurrence dependent on bottleneck location and traffic demand

The probability of the first pattern occurrence (e.g., from Fig. 42 at a typical bottleneck) is in itself distributed over time. After the traffic breakdown, the further development of the congested region is more deterministic than the first phase transition (occurring from free to synchronized flow in Fig. 42 at approx. 15 km and 14:20 h).

Matching with Current Traffic Situation for Prediction

If a structured database with sets of time series of traffic variables exists, the current traffic situation can be used as criterion for choice of the “best fit” prediction (Fig. 43). Measurements in the vehicle or at the traffic center support the search in the archive: a spatiotemporal criterion determines the prediction output. The choice of the “best fit” from the historical archive should classify a “reference” pattern from an anomaly. A time-shift component has to be established to ensure that, for example, a morning peak 10 min later is not considered as an anomaly, but still a regular

pattern (Kerner 2004; Venkatanarayana et al. 2005). Because of the features of the traffic process, the acceptable tolerance defining similarity among patterns has to integrate spatiotemporal distances. “Best fit” time series must therefore be found in the environment of the specific location, i.e., neighboring detectors or road sections, as well as with time shifts of the current prediction time.

Applications of Traffic Prediction

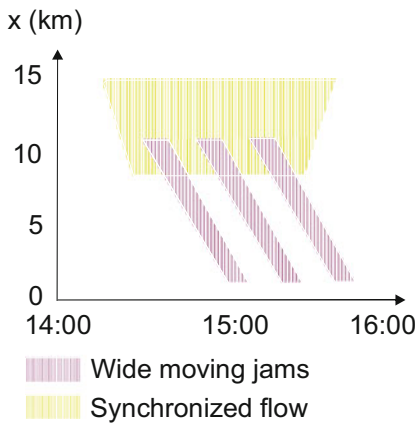
Currently known applications of traffic prediction are split into several areas. They could be implemented on the infrastructure or the vehicle

side. The internet offers a widely used medium for distributing traffic prediction to the user without individualizing the prediction information which is dependent on current position and time. The online application of FOTO and ASDA models offers traffic prediction information which can be used in the traffic control center for traffic management purposes. An application concept for prediction usage in vehicles closes this short overview.

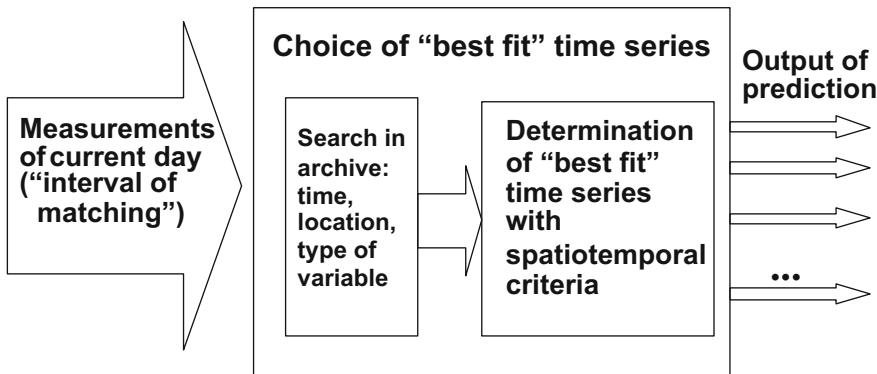
Traffic Control Centers

For the control of freeway traffic, traffic control centers have some instrumentation like speed advisory, lane blockings, collective route choice, etc. To decide the best possible control strategy, a traffic prediction is relevant information. In the control center of the German federal state of Hessen, online traffic prediction based on FOTO and ASDA models has been installed and used.

The online application – called FOTOWin integrating both FOTO and ASDA models – offers the possibility to predict the movement of a wide moving jam (see above section “Prediction of Propagation of the Fronts of Wide Moving Jam”). In case of wide moving jam recognition, the online system predicts the subsequent propagation of the related wide moving jam. This information could be very useful for vehicle applications of local danger warnings: the sudden traffic breakdown of vehicle speed could be predicted, if such a wide moving jam approaches the individual vehicle.



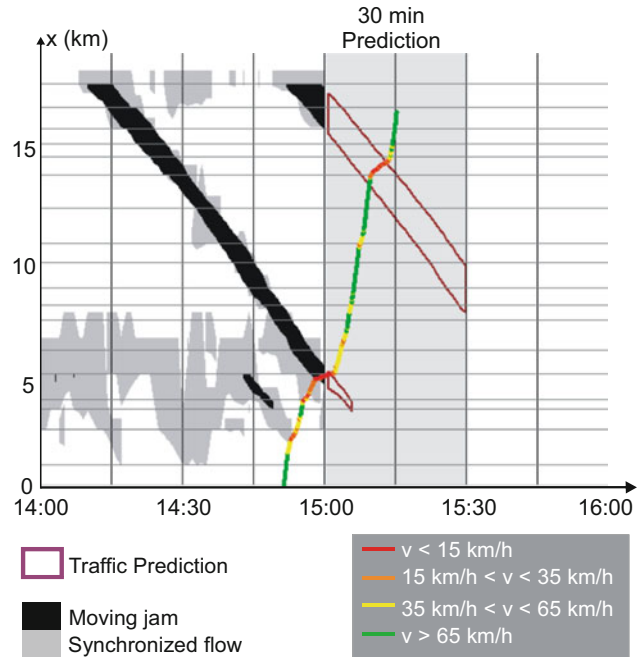
Traffic Prediction of Congested Patterns, Fig. 42 Time series of congested traffic pattern database (general pattern)



Traffic Prediction of Congested Patterns, Fig. 43 Matching of time series with current measurements

Traffic Prediction of Congested Patterns,

Fig. 44 Empirical wide moving jam short-term prediction with FOTO and ASDA models for a vehicle (colored trajectory)



An empirical example is illustrated in Fig. 44. At 15:00 h the future positions of a wide moving jam have been predicted. The test vehicle has met the upstream front at 14 km at about 15:08 h. A local danger warning of a speed breakdown could have been given about 8 min and 9 km in advance. Inside the vehicle, this could be used for driver information, safety, and comfort functions. If this information could be given with enough quality in larger road networks, the use for vehicle assistance systems like adaptive cruise control becomes technically feasible.

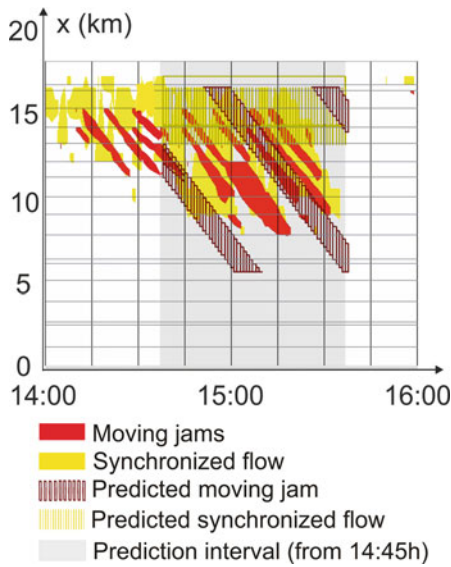
The online FOTOWin system in Hessen is able to “match” the current traffic pattern with a pattern database. Therefore, the infrastructure gets information on the subsequent development of the congested traffic pattern. This is relevant for decision support for the current traffic management strategy.

One empirical result of a short-term pattern prediction with matching is represented in the following space-time diagram (Fig. 45). The horizontal lines represent the positions of the detectors on the freeway. The light grey area denotes the prediction period with red and yellow vertical lines signifying wide moving jams and synchronized flow, respectively. The related matched traffic database

elements from the pattern database are also shown. Together with the prediction from 14:38 h, the reality is shown, which has been reconstructed later.

The online FOTOWin system has been installed recently for North-Rhine Westphalia freeways (Palmer and Rehborn 2007). In principle, the traffic control center located there has similar detector loop data available as in Hessen. These raw traffic data are transferred to WDR, the major public radio broadcasting station from North-Rhine Westphalia in Cologne, who offers traffic messages to the end customer (e.g., radio listener or driver) via broadcast channel RDS (Radio Data System). The application FOTOWin covers a part of the whole freeway network with 1900 km of freeway and more than 1000 double loop detectors (Fig. 46).

As an empirical example, a large congested pattern on freeway A45-North is illustrated in Fig. 47. It is clearly visible that such a pattern is very similar to congested traffic patterns in Hessen (see Fig. 30): emergence of wide moving jams and their stable propagation through many other bottlenecks does eventually occur on every freeway with bottlenecks.



Traffic Prediction of Congested Patterns, Fig. 45 Traffic pattern prediction, freeway A3-North, 06 June 2006, prediction at 14:38 h



Traffic Prediction of Congested Patterns, Fig. 46 Freeway road network observed by FOTOWin in North-Rhine Westphalia (Palmer and Rehborn 2007)

In this example, for almost 7 h on the 17 km stretch a large number of propagating wide moving jams have emerged. At the locations approximately at 10 and 16 km phase transitions from

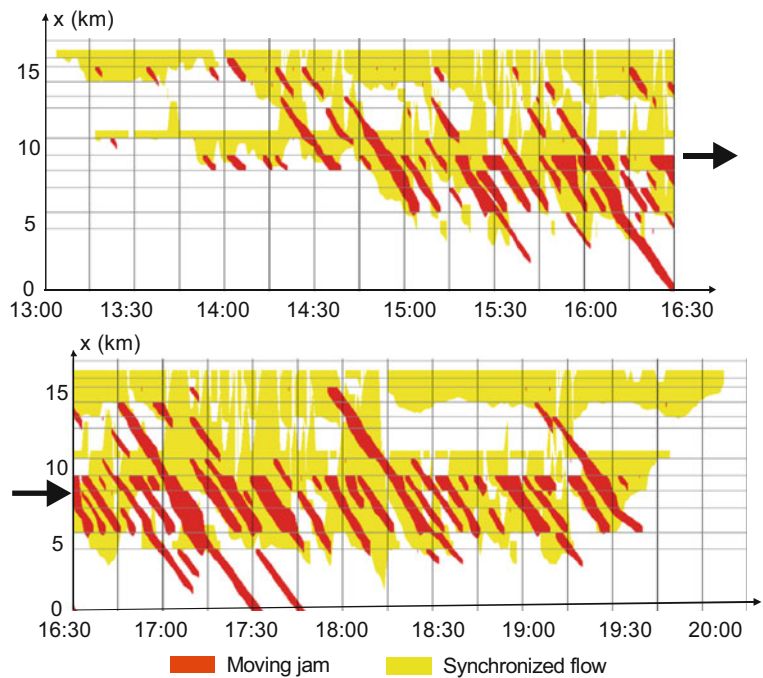
free to synchronized flow occur at different time moments. Later on, wide moving jams propagate from the downstream bottleneck at approximately 16 km to the upstream bottleneck in the vicinity of 10 km: the occurring empirical expanded pattern illustrates the complex interactions of congested traffic patterns which can be explained based on Kerner three-phase traffic theory. Because of the dense freeway network in North-Rhine Westphalia, there are many possible bottlenecks on the freeways due to a large variety of on- and off-ramps with the occurrence of EPs.

In-Vehicle Traffic Prediction

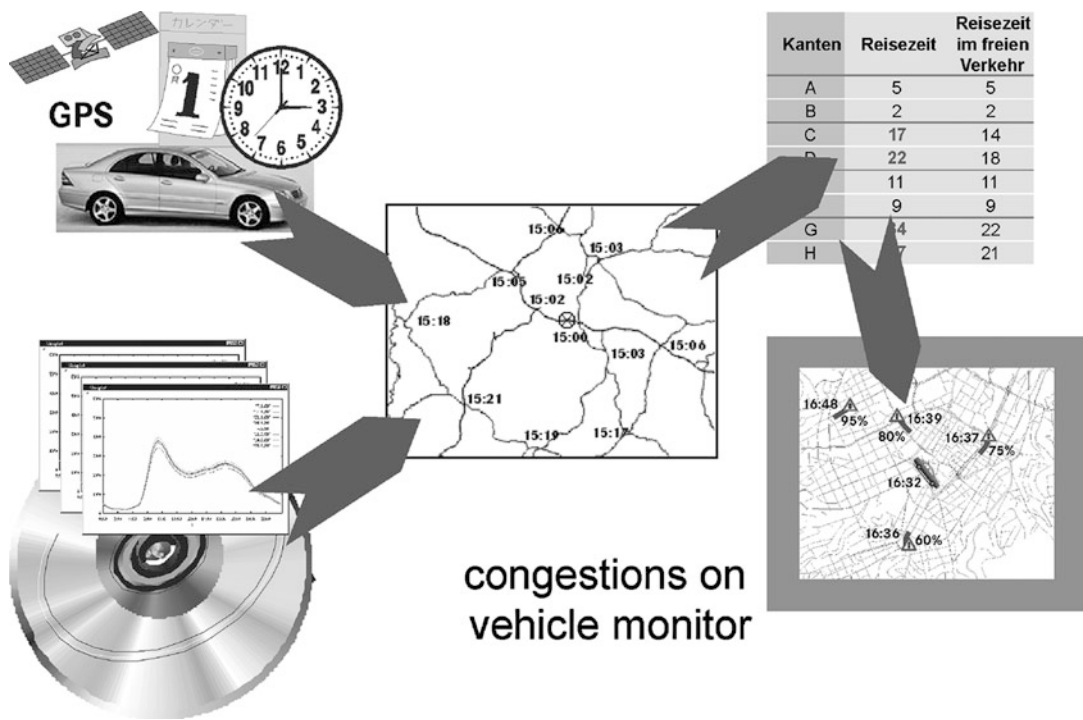
The concept of autonomous traffic prediction in a vehicle is presented in Kerner et al. (2005). Figure 48 illustrates the approach: a vehicle has a position unit, a digital map, and the current time. Additionally, the vehicle has an on board memory unit for historical travel time series (e.g., stored on CD/DVD). The linkage of the map information with stored traffic congestion information leads to an adequate choice of travel time prediction associated with the current time and the specific vehicle location. Subsequently, the driver can see this traffic prediction indicated in a vehicle monitor. Furthermore, the traffic prediction is useful for dynamic route guidance in navigation systems (e.g., Chen et al. 2007). Results of traffic prediction can also be presented in different visualization forms (Kerner et al. 2005).

A possible extension of the above concept is the integration of historical spatiotemporal patterns (Kerner 2004) to also make short-range traffic prediction available (Fig. 49). To reach this goal, a historical spatiotemporal pattern database should be stored in the vehicle, in addition to historical travel times (or flow rates). In this case, based on the current vehicle location and speed measured by the vehicle, an appropriate choice of a spatiotemporal pattern can be made in the vehicle, which the vehicle should meet in the near future (see vehicle trajectory in Fig. 49, bottom, right).

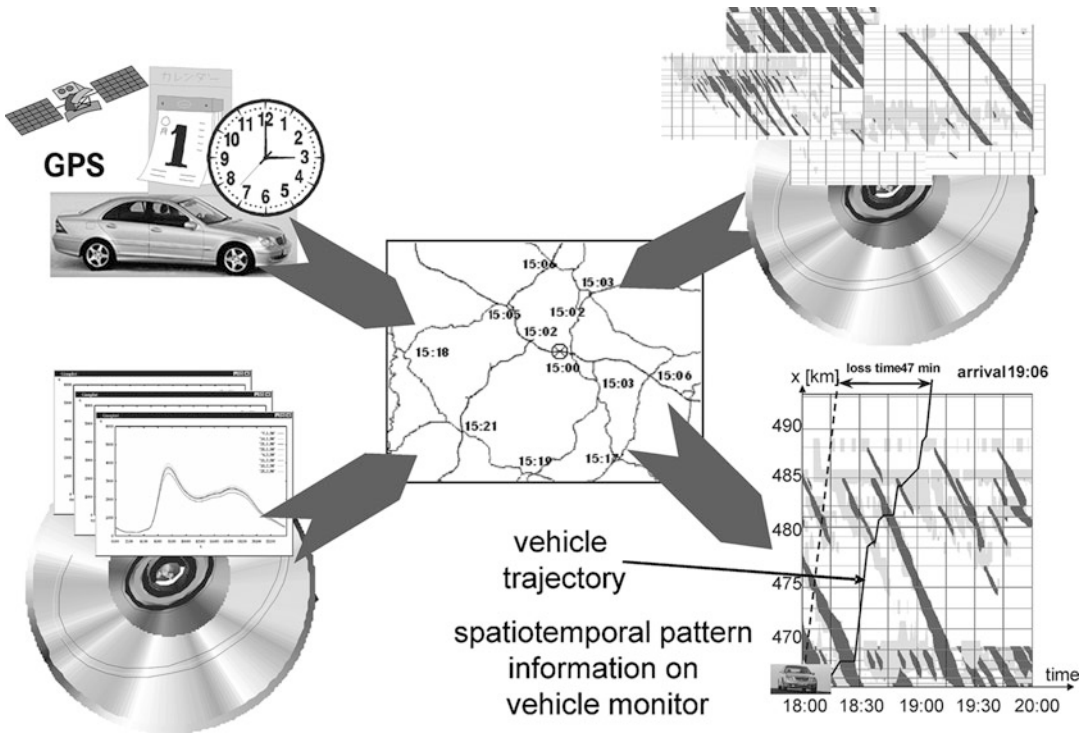
As it is shown (e.g., Ding et al. 2002), traffic is highly regular with recurring and typical characteristics. Therefore, the historical patterns are relatively stable within proposed update cycles of



Traffic Prediction of Congested Patterns, Fig. 47 Congested traffic pattern reconstructed by FOTO and ASDA models: space-time diagram from freeway A45-North in North-Rhine Westphalia, 31 August 2007 (Palmer and Rehborn 2007)



Traffic Prediction of Congested Patterns, Fig. 48 Concept of autonomous traffic prediction (Kerner et al. 2005)



Traffic Prediction of Congested Patterns, Fig. 49 Concept of autonomous traffic prediction with use of spatiotemporal pattern database in vehicle (Kerner et al. 2005)

1–2 years because of long-term changes in traffic demand and infrastructure: the pattern database in the vehicles could be relatively steady. Communication channels into the vehicle could update the spatiotemporal pattern database.

Future Directions

It must be stated that in the field of traffic prediction, a large number of different concepts are coexisting. The proof of reliability especially in times of congested traffic is a still open question to the traffic prediction approaches. Traffic predictions in online information systems and used for collective traffic management or individual navigation system in established applications are still in its infancies because of the heterogeneous and often unreliable prediction methods. Therefore, quality definitions and measurement techniques are a key issue for the future evaluation of traffic prediction approaches.

One of the missing elements is a quality definition for precise traffic predictions, because the step beforehand – an area covering precise measuring of the current traffic situation – is still not realized. The measurements of the current situation could be enhanced by using the vehicle as a moving sensor (FCD, floating car data) or using any mobile device in the vehicle (FPD, flow phone data). Both technological ideas have been understood conceptually and are introduced in the market. A definition of prediction quality tested in a laboratory environment should be based on deviation of the predicted values from real measured and archived travel times of the congested time periods.

Microscopic simulation is a very helpful instrument to predict consequences of both traffic control as well as individual traffic assistance systems (e.g., automatic cruise control) before introducing them into traffic control center or into the vehicles. Especially microscopic modeling of vehicle behavior in the case of automatic

vehicle assistance systems and their influence of the global traffic system is a complex field of vehicle dynamics. A quality assured network wide traffic prediction with microscopic simulation analogous to weather prediction is still a task of the future.

The traffic pattern database could be enhanced for planned roadwork and accidents which both show characteristic predictable features. The traffic pattern database then has the characteristic content of long-term time series information.

A short-term precise prediction could be used as input for driver assistance systems in vehicles, e.g., for an adapted automatic speed control of the vehicle in the case of approaching congestion.

Traffic simulation in good accordance with the reality on the roads will be a valuable instrument for network traffic predictions. Benchmarking criteria for evaluation of model results should be defined to have quality measures for their application.

One relevant further issue is the enhancement and combination of long-term time series with time series of breakdown probabilities. Anticipation effects will arise with further distribution of dynamic traffic information: people will react on the information in different ways adapting their routing decisions. As consequence, the problem of dynamic traffic assignment under more and more dynamic current traffic information increases: how should the long-term traffic prediction information be given to have an efficient use of the available network resources and not the problem of emerging new congestion in regions where people use roads trying to avoid the already congested regions.

Looking on the network perspective, the effects of prediction on the driver behavior must be taken into account if larger information penetration rates are realized. In case of more available information, the dynamic traffic assignment problem arises, i.e., which part of the driver collective should be advised on which route, finding a system optimum for the road network. The most recent book by Kerner (2017) describes a “breakdown minimization” principle, where traffic congestion is controlled and reduced by minimizing the breakdown probability at each of the network bottlenecks.

Bibliography

Primary Literature

Patents information at www.depatistnet.de

- Abdulhai B, Porwal H, Recker W (1999) Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks. In: Proceedings 78th annual meeting Transportation Research Board. National Academies Press, Washington, DC
- Acha-Daza JA, Hall FL (1993) A graphical comparison of the predictions for speed given by catastrophe theory and some classic models. *Transp Res Rec* 1398: 119–124
- Ahmed MS, Cook AR (1979) Analysis of freeway traffic time-series data by using Box-Jenkins techniques. *Transp Res Rec* 722:1–9
- Arem BV, Kirby HR, Van Der Vlist MJM, Whittaker JC (1997) Recent advances and applications in the field of short-term traffic forecasting. *Int J Forecast* 13:1–12
- Becker M, Fastenrath U (1998) Method for transmitting local data and measurement data from a terminal, including a telematic terminal, to a central traffic control unit. German Patent Publication DE 197 55 875 A1, USA: US6426709B1
- Ben-Akiva M, Cuneo D, Hasan M, Jha M, Yang Q (2003) Evaluation of freeway control using a microscopic simulation laboratory. *Transp Res C Emerg Technol* 11(1):29–50
- Boker G, Lunze J (2001) State estimation in freeway traffic with floating car data. *Automatisierungstechnik* 49(11): 497–504
- Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control. Holden-Day, San Francisco
- Brockwell PJ, Davis RA (2002) Introduction to time series and forecasting, 2nd edn. Springer, New York
- Burrus CS, Gopinath RA, Guo HT (1998) Introduction to wavelets and wavelet transforms: a primer. Prentice Hall, Upper Saddle River
- Cetin M, Comert G (2006) Short-term traffic flow prediction with regime switching models. *Transp Res Rec* 1965:23–31
- Chatfield C (2001) Time-series forecasting. Chapman & Hall/CRC, London
- Chen M, Chien SIJ (2001) Dynamic freeway travel-time prediction with probe vehicle data – link based versus path based. *Transp Res Rec* 1768:157–161
- Chen H, Grant-Muller S (2001) Use of sequential learning for short-term traffic flow forecasting. *Transp Res C* 9:319–336
- Chen H, Grant-Muller S, Mussone L, Montgomery F (2001) A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. *Neural Comput Appl* 10:277–286
- Chen Y, Bell MGH, Bogenberger K (2007) Reliable multi-path planning and dynamic adaptation for a centralized road navigation system. *IEEE Trans ITS* 8(1):14–20
- Chickering DM, Heckerman D, Meek C (1997) A Bayesian approach to learning Bayesian networks with local structure. In: Proceedings 13th conference on

- uncertainty in artificial intelligence, Rhode Island, pp 80–89
- Chien SIJ, Kuchipudi CM (2003) Dynamic travel time prediction with real-time and historic data. *J Transp Eng ASCE* 129(6):608–616
- Chrobok R, Wahle J, Schreckenberg M (2001) Traffic forecast using simulations of large scale networks. In: Stone B, Conroy P, Broggi A (eds) 4th international IEEE conference on intelligent transportation systems. IEEE, Oakland, pp 434–439
- Cremer M (1979) Traffic flow on freeways. Springer, Berlin. (in German)
- D'Angelo MP, Al-Deek HM, Wang MC (1999) Travel-time prediction for freeway corridors. *Transp Res Rec* 1676:184–191
- Daganzo CF (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B* 28(4): 269–287
- Daganzo CF (1995) The cell transmission model, part II: network traffic. *Transp Res B* 29(2):79–93
- Daganzo CF (1997) Fundamentals of transportation and traffic operations. Elsevier Science, Oxford, UK
- Daganzo CF (1999) The lagged cell-transmission model. In: Ceder A (ed) Proceedings of the 14th international symposium on transportation and traffic theory. Elsevier Science, Jerusalem, Israel, pp 81–104
- Davis GA, Nihan NL (1991) Nonparametric regression and short-term freeway traffic forecasting. *J Transp Eng* 117(2):178
- de Rham C, Lange R (2000) Short term forecast and evaluation for intelligent VMS settings. In: Proceedings of the 7th world congress on ITS, Torino
- Dharia A, Adeli H (2003) Neural network model for rapid forecasting of freeway link travel time. *Eng Appl Artif Intell* 16(7–8):617–613
- Dia H (2001) An object oriented neural network approach to short term traffic forecasting. *Eur J Oper Res* 131:253–261
- Ding A, Zhao X, Jiao L (2002) Traffic flow time series prediction based on statistics learning theory. In: IEEE 5th international conference on intelligent transportation systems, Singapore, pp 727–730
- Dion F, Rakha H, Kang YS (2004) Comparison of delay estimates at under-saturated and 38 over-saturated pre-timed signalized intersections. *Transp Res B Methodol* 38(2):99–122
- Disbro JE, Frame M (1989) Traffic flow theory and chaotic behaviour. *Transp Res Rec* 1225:109–125
- Dougherty M (1995) A review of neural networks applied to transport. *Transp Res C* 3(4):247–260
- Edie LC, Foote RS (1960) Effect of shock waves on tunnel traffic flow. In: Highway Research Board – proceedings 39. National Research Council, Washington, DC, pp 492–505
- Fallah-Tafti M (2001) The application of artificial neural networks to anticipate the average journey time of traffic in the vicinity of merges. *Knowl-Based Syst* 14:203–211
- Fastenrath U (1998) Method for determining traffic data and traffic information exchange. German Patent Publication DE 197 37 440 A1, USA: US 6329932B1
- Fuller WA (1996) Introduction to statistical time series, 2nd edn. Wiley, New York
- Gartner NH, Stamatiadis C (2008) Optimization and control of urban traffic networks. This encyclopedia. Springer, New York
- Gazis D, Knapp C (1971) Online estimation of traffic densities from time series of traffic and speed data. *Transp Sci* 5:283–301
- Geroliminis N, Skabardonis A (2011) Identification and analysis of queue spillovers in city street networks. *IEEE Trans Intell Transp Syst* 12(4):1107–1115
- Gipps PGA (1981) A behavioural car-following model for computer simulation. *Transp Res B* 15:105–111
- Grenander U (1996) Elements of pattern theory. Johns Hopkins University Press, Baltimore
- Hamed MM, Al-Masaeid HR, Bani Said ZM (1995) Short-term prediction of traffic volume in urban arterials. *J Transp Eng* 121(3):249–254
- Haykin S (1999) Neural networks: a comprehensive foundation. Prentice Hall, Upper Saddle River
- Head LK (1995) Event-based short-term traffic flow prediction model. *Transp Res Rec* 1510:45–52
- Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley, Reading
- Heidemann D, Wimber P (1982) Types of traffic flow rate time series based on clustering methods. In: Straßenverkehrszählungen, vol 26. BAST, Germany
- Helbing D (1997) Traffic dynamics: new modelling concepts in physics. Springer, Berlin/Heidelberg. (in German)
- Hemmerle P, Koller M, Rehborn H, Kerner BS, Schreckenberg M (2016a) Fuel consumption in empirical synchronised flow in urban traffic. *IET Intell Transp Syst* 10(2):122–129
- Hemmerle P, Koller M, Hermanns G, Schreckenberg M, Rehborn H, Kerner BS (2016b) Impact of synchronised flow in oversaturated city traffic on energy efficiency of conventional and electrical vehicles. In: Knoop V, Daamen W (eds) Traffic and granular flow '15. Springer, Cham
- Hemmerle P, Koller M, Hermanns G, Rehborn H, Kerner BS, Schreckenberg M (2016c) Impact of synchronised flow in oversaturated city traffic on energy efficiency of conventional and electrical vehicles. *Collect Dyn* 1:1–27
- Hermanns G, Hemmerle P, Rehborn H, Koller M, Kerner BS, Schreckenberg M (2015) Microscopic simulation of synchronized flow in oversaturated city traffic: effect of drivers speed adaptation. *Transp Res Rec J Transp Res Board* 2490:47–55
- Hermanns G, Hemmerle P, Rehborn H, Kerner BS, Schreckenberg M (2016) Microscopic simulations of oversaturated city traffic: features of synchronised flow patterns. In: Knoop V, Daamen W (eds) Traffic and granular flow '15. Springer, Cham
- Highway Capacity Manual 2000 (2000) Transportation Research Board. National Research Council, Washington, DC
- Horvitz E, Apacible J, Sarin R, Liao L (2005) Prediction, expectation, and surprise: methods, designs, and study of a deployed traffic forecasting service. In: Proceedings of

- the conference on uncertainty and artificial intelligence 2005. AUAI Press, Edinburgh, Scotland
- Hoyer R, Chrobok R, Feldges M, Folkerts G, Friedrich B, Huber W, Kates R, Kemper C, Kirschfink H, Lange R, Listl G, Mathias P, Offermann F, Pinkofsky L, Rehborn H, Schlichting B, Stieler P, Thiemann O, Vortisch P (2003) Advice for data completion and data aggregation in traffic management applications. *Hinweispapier der Forschungsgesellschaft für Straßen- und Verkehrswesen, FGSV-Papier*, vol 382. (in German)
- Huang SH, Ran B (2003) An application of neural network on traffic speed prediction under adverse weather condition. In: 82nd TRB annual meeting. National Academies Press, Washington, DC
- Huiskens G, Van Berkum EC (2003) A comparative analysis of short-range travel time prediction methods. In: 82nd TRB annual meeting Transportation Research Board. National Academies Press, Washington, DC
- Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT – a traffic responsive method of coordinating signals. TRRL report no. LR1014, Transport and Road Research Laboratory, Crowthorne
- Innema S (2001) Short term prediction of highway travel time using MLP neural networks. In: 8th world congress on intelligent transportation systems, Sydney, pp 1–12
- Ishak S, Al-Deek H (2002) Performance evaluation of short term time series traffic prediction model. *J Transp Eng ASCE* 128(6):490–498
- Ishak S, Alecsandru C (2004) Optimizing traffic prediction performance of neural networks under various topological, input, and traffic condition settings. *J Transp Eng* 130:452–465
- Jiang X, Adeli H (2004) Wavelet packet-autocorrelation function method for traffic flow pattern analysis. *Comput Aided Civ Infrastruct Eng* 19:324–337
- Kaumann O, Froese K, Chrobok R, Wahle J, Neubert L, Schreckenberg M (2000) Online simulation of the freeway network of NRW. In: Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) *Traffic and granular flow '99*. Springer, Berlin Heidelberg, pp 351–356
- Kaysi I, Ben-Akiva M, Koutsopoulos H (1993) An integrated approach to vehicle routing and congestion prediction for real-time driver guidance. *Transp Res Rec* 1408:66–74
- Kerner BS (1998) Experimental features of self-organization in traffic flow. *Phys Rev Lett* 81:3797
- Kerner BS (1999a) Traffic prediction method for road network with traffic controlled network nodes. German Patent DE 199 40 957 C2. (in German)
- Kerner BS (1999b) Method for monitoring the condition of traffic for a traffic network comprising effective narrow points. German Patent DE 199 44 075 C2, USA Patent: US6813555B1; Japan Patent: JP 2002117481
- Kerner BS (1999c) Congested traffic flow: observations and theory. *Transp Res Rec* 1678:160–167
- Kerner BS (1999d) Theory of congested traffic flow: self-organization without bottlenecks. In: 14th international symposium on transportation and traffic theory. Jerusalem, Israel, pp 147–171
- Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
- Kerner BS (2004) *The physics of traffic*. Springer, Berlin/New York
- Kerner BS (2007) On-ramp metering based on three-phase traffic theory. *Traffic Eng Control* 48(1):28–35
- Kerner BS (2008) Modelling approaches to traffic congestion. This encyclopaedia. Springer, New York
- Kerner BS (2009) *Introduction to modern traffic flow theory and control*. Springer, Berlin/New York
- Kerner BS (2014) Cumulated vehicle acceleration. *Traffic Eng Control* 55(4):139–141
- Kerner BS (2017) Breakdown in traffic networks: fundamentals of transportation science. Springer, Berlin
- Kerner BS, Herrtwich RGH (2001) Traffic forecasting. *Automatisierungstechnik* 49:505–511
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35(3):L31–L43
- Kerner BS, Klenov SL (2003) A microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68(3):036130
- Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A Math Gen* 39:1775–1809
- Kerner BS, Konhäuser P (1994) Structure and parameters of clusters in traffic flow. *Phys Rev E* 50(1):54
- Kerner BS, Rehborn H (1996a) Experimental properties of complexity in traffic flow. *Phys Rev E* 53:R4257
- Kerner BS, Rehborn H (1996b) Experimental features and characteristics of traffic jams. *Phys Rev E* 53:1297
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. *Phys Rev Lett* 79:4030
- Kerner BS, Rehborn H (1998) Traffic surveillance method and vehicle flow control in a road network. German Patent Publication DE 198 35 979 A1, USA Patent: US 6587779B1
- Kerner BS, Rehborn H, Kirschfink H (1998) Method for the automatic monitoring of traffic including the analysis of back-up dynamics. German Patent DE 196 47 127 C2, Dutch Patent: NL1007521C, USA Patent US 5861820
- Kerner BS, Aleksic M, Denneler U (1999) Traffic condition supervision in traffic network, undertaking inquiry of current position and/or prognosis of future position of flank between area of free traffic and area of synchronized traffic continuously. German Patent DE 199 44 077 C1
- Kerner BS, Rehborn H, Aleksic M, Haug A (2004) Recognition and tracing of spatial-temporal congested traffic patterns on freeways. *Transp Res C* 12:369–400
- Kerner BS, Rehborn H, Haug A, Aleksic M (2005) Traffic prediction in vehicles. In: *Proceedings 8th IEEE conference on intelligent transportation systems*, Vienna, pp 251–256

- Kerner BS, Rehborn H, Palmer J, Klenov SL (2011) Using probe vehicle data to generate jam warning messages. *Traffic Eng Control* 3:141–148
- Kerner BS, Klenov SL, Hermanns G, Hemmerle P, Rehborn H, Schreckenberg M (2013a) Synchronized flow in oversaturated city traffic. *Phys Rev E* 88(5):054801
- Kerner BS, Rehborn H, Schäfer RP, Klenov SL, Palmer J, Lorkowski S, Witte N (2013b) Traffic dynamics in empirical probe vehicle data studied with three-phase theory: spatiotemporal reconstruction of traffic phases and generation of jam warning messages. *Phys A Stat Mech Appl* 392(1):221–251
- Kerner BS, Hemmerle P, Koller M, Hermanns G, Klenov SL, Rehborn H, Schreckenberg M (2014) Empirical synchronized flow in oversaturated city traffic. *Phys Rev E* 90(3):032810
- Kirby HR, Watson SM, Dougherty MS (1997) Should we use neural networks or statistical models for short-term motorway traffic forecasting? *Int J Forecast* 13:43–50
- Kirschfink H (1999) Collective traffic control in motorways. Tutorial at the 11th EURO-mini conference on AI in transportation systems and science, Helsinki
- Kirschfink H, Hernández J, Boero M (2000) Intelligent traffic management models. In: *Proceedings of the European symposium on intelligent techniques (ESIT)*. Helsinki, Finland
- Kisgyorgy L, Rilett LR (2002) Travel time prediction by advanced neural network. *Period Polytech Ser Civ Eng* 46(1):15–32
- Kitamura K, Kuwahara M (eds) (2005) *Simulation approaches in transportation analysis: recent advances and challenges*. Operations research/computer science interfaces series, vol 31. Springer, US
- Kniss HC (2000) Evaluation of ASDA/FOTO in traffic control centre Hessen (internal report, in German)
- Koller M, Hemmerle P, Rehborn H, Hermanns G, Kerner BS, Schreckenberg M (2014) Increased consumption in synchronized flow in oversaturated city traffic. In: 10th ITS European congress, Helsinki, proceedings
- Koller M, Hemmerle P, Rehborn H, Kerner BS, Kaufmann S (2016) Traffic phase dependent fuel consumption. In: Knoop V, Daamen W (eds) *Traffic and granular flow '15*. Springer, Cham
- Koshi M, Iwasaki M, Ohkura I (1983) Some findings and an overview on vehicular flow characteristics. In: *Proceedings 8th international symposium on transportation and traffic theory*. Toronto, Canada, p 403
- Kuchipudi CM, Chien SIJ (2003) Development of a hybrid model for dynamic travel time prediction. In: 82nd annual meeting Transportation Research Board, Washington, DC
- Kwon J, Coifman B, Bickel P (2000) Day-to-day travel-time trends and travel-time prediction from loop-detector data. *Transp Res Rec* 1717:120–129
- Lan CJ, Miaou SP (1999) Real-time prediction of traffic flows using dynamic generalized linear models. *Transp Res Rec* 1678:168–178
- Lee S, Kim D, Kim J, Cho B (1998) Comparison of models for predicting short-term travel speeds. In: 5th world congress on intelligent transport systems, Seoul
- Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
- Lieu HC (2000) Traffic estimation and prediction system. *Transp Res News* 208:3–6
- Lindveld CDR, Thijs R, Bovy PHL, Van der Zijpp NJ (2000) Evaluation of online travel time estimators and predictors. *Transp Res Rec* 1719:45–53
- Lingras P, Sharma S, Zhong M (2002) Prediction of recreational travel using genetically designed regression and time-delay neural network models. *Transp Res Rec* 1805:16–24
- Lu J (1990) Prediction of traffic flow by an adaptive prediction system. *Transp Res Rec* 1287:13–20
- Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419:1–64
- Matsui H, Fujita M (1998) Travel time prediction for freeway traffic information by neural network driven fuzzy reasoning. In: Himanen V, Nijkamp P, Reggiani A, Raito J (eds) *Neural networks in transport applications*. Ashgate Publishers, Burlington, pp 355–364
- May AD (1990) *Traffic flow fundamentals*. Prentice Hall, Englewood Cliffs
- Middelham F (2001) Predictability: some thoughts on modelling. *Futur Gener Comput Syst* 17(5):627–636
- Miyata S, Noda M, Usami T (1995) STREA. In: *Proceedings of the 2nd world congress on intelligent transport systems*, Yokohama, vol 1, pp 289–297
- Moorthy CK, Ratcliffe BG (1998). Short term traffic forecasting using time series methods. *Transp Plan Technol* 12(1):45–56
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I Fr* 2:2221–2229
- Nam DH, Drew DR (1996) Traffic dynamics: method for estimating freeway travel times in real time from flow measurements. *J Transp Eng* 122(3):186–191
- Nanthawichit C, Nakatsuji T, Suzuki H (2003) Application of probe vehicle data for real-time traffic state estimation and short term travel time prediction on a freeway. In: *Proceedings 82nd annual meeting Transportation Research Board*, Washington, DC
- Newell GF (1965) Approximation methods for queues with application to the fixed-cycle traffic light. *SIAM Rev* 7(2):223–240
- Newell GF (1982) *Applications of queuing theory*. Chapman & Hall, London
- Nicholson H, Swann CD (1974) The prediction of traffic flow volumes based on spectral analysis. *Transp Res* 8:533–538
- Nihan NL, Holmesland KO (1980) Use of the Box and Jenkins time series technique in traffic forecasting. *Transportation* 9:125–14372
- Nikovski D, Nishiuma N, Goto Y, Kumazawa H (2005) Univariate short-term prediction of road travel times. In: *International IEEE conference on intelligent transportation systems (ITSC)*, Vienna

- Ober-Sundermeier A, Zackor H (2001) Prediction of congestion due to road works on freeways. In: Proceedings IEEE intelligent transportation systems, Oakland, pp 240–244
- Oda T (1990) An algorithm for prediction of travel time using vehicle sensor data. In: IEEE 3rd international conference on road traffic control. London, pp 40–44
- Oh C, Ritchie SG, Oh JS (2005) Exploring the relationship between data aggregation and predictability toward providing better predictive traffic information. *Transp Res Rec* 1935:28–36
- Ohba Y, Koyama T, Shimada S (1997) Online learning type of travelling time prediction model in expressway. In: IEEE conference on intelligent transport systems, Boston, pp 350–355
- Okutani I, Stephanedes YI (1984) Dynamic prediction of traffic volume through Kalman filtering theory. *Transp Res B* 18B(1):1–11
- Palmer J, Rehborn H (2008) ASDA/FOTO based on Kerner's three-phase traffic theory in North-Rhine Westfalia. *Straßenverkehrstechnik*. (in German) 8:463–470
- Pancratz A (1991) Forecasting with dynamic regression models. Wiley-InterScience, New York
- Papageorgiou M (1983) Application of automatic control concepts in traffic flow modelling and control. Springer, Berlin/New York
- Park B, Messer CJ, Urbanik T II (1998) Short term traffic volume forecasting using radial basis function neural network. *Transp Res Rec* 1651:39–47
- Park DJ, Rilett LR, Han G (1999) Spectral basis neural networks for real-time travel time forecasting. *J Transp Eng* 125(6):515–523
- Petty KF, Bickel P, Ostland M, Rice J, Schoenberg F, Jiang J, Ritov Y (1998) Accurate estimation of travel times from single loop detectors. *Transp Res A* 32(1):1–17
- Pinkofsky L (2002) Types of time series. In: Verkehrsentwicklung auf Bundesfernstraßen 2002. Bericht der Bundesanstalt für Straßenwesen, Reihe Verkehrstechnik, vol V99. Bergisch Gladbach. Bundesanstalt für Straßenwesen (BASt) (in German)
- Qiao F, Wang X, Yu L (2003) Optimizing aggregation level for ITS data based on wavelet decomposition. In: Proceedings 82nd annual meeting Transportation Research Board. National Academies Press, Washington, DC
- Rakha H, Crowther B (2003) Comparison and calibration of FRESIM and INTEGRATION steady-state car-following behaviour. *Transp Res A* 37:1–27
- Ran R, Boyce D (1996) Modelling dynamic transportation networks. Springer, Berlin
- Rehborn H, Haug A, Aleksic M, Kerner BS, Fastenrath U (2002) Statistical analysis of traffic message archives as decision support for road construction up to traffic management. *Straßenverkehrstechnik* 9:478–485. (in German)
- Rehborn H, Haug A, Kerner BS, Aleksic M, Fastenrath U (2003) Floating car data and methods for recognition and tracking of spatiotemporal traffic patterns. *Straßenverkehrstechnik* 9:461–468. (in German)
- Rehborn H, Klenov SL, Palmer J (2011) An empirical study of common traffic congestion features based on traffic data measured in the USA, the UK, and Germany. *Phys A Stat Mech Appl* 390(23):4466–4485
- Rice J, Van Zwet E (2001) A simple and effective method for predicting travel times on freeways. In: Proceedings of the IEEE conference on intelligent transportation systems, Oakland, pp 227–232
- Riegelhuth G, Kirschfink H (2003) Management with decision support of road works for traffic flow optimization on freeways. In: Proceedings of ITS world congress, paper no. 2255T
- Rilett LR, Park D (2001) Direct forecasting of freeway corridor travel times using spectral basis neural networks. *Transp Res Rec* 1752:140–147
- Robertson DI (1969) TRANSYT: a traffic network study tool. TRRL report no. LR 253, Transportation and Road Research Laboratory, Crowthorne
- Rumelhart DE, McClelland JL (1986) Parallel distributed processing: exploration in the microstructure of cognition. MIT Press, Cambridge, MA
- Schönhof M, Helbing D (2007) Empirical features of congested traffic states and their implications for traffic modeling. *Transp Sci* 41(2):135–166
- Schrader CC, Kornhauser AL, Friesen LM (2004) Using historical information in forecasting travel times. In: 82nd annual meeting Transportation Research Board. National Academies Press, Washington, DC
- Smith BL, Demetsky MJ (1994) Short term traffic flow prediction: neural network approach. *Transp Res Rec* 1453:98–104
- Smith BL, Demetsky MJ (1997) Traffic flow forecasting: comparison of modeling approaches. *J Transp Eng* 123(4):261–266
- Smith BL, Oswald KR (2003) Meeting real-time traffic flow forecasting requirements with imprecise computations. *Comput Aided Civ Infrastruct Eng* 18:201–213
- Smith BL, Williams BM, Oswald KR (2002) Comparison of parametric and non-parametric models for traffic flow forecasting. *Transp Res C* 10(4):303–321
- Stathopoulos A, Karlaftis MG (2003) A multivariate state-space approach for urban traffic flow modeling and prediction. *Transp Res C* 11:121–135
- Sun H, Liu HX, Xiao H, He RR, Ran B (2003) Short-term traffic forecasting using the local linear regression model. *J Transp Res Board* 1836:143–150
- Sun H, Xiao HX, Yang F, Ran B, Tao Y, Oh Y (2004) Wavelet preprocessing for local linear traffic prediction. In: 83rd Transportation Research Board annual meeting, Washington, DC
- Teng H, Qi Y (2003) Application of wavelet technique to freeway incident detection. *Transp Res C* 11(3–4):289
- Traffic Flow Theory 2006 (2006) Monograph with 22 papers on the subject of traffic flow theory. Transportation Research Record 1965. Transportation Research Board, Washington

- Treiterer J (1975) Investigations of traffic dynamics by aerial photogrammetry. Ohio State University Technical Report PB 246 094, Columbus
- Van der Voort M, Dougherty M, Watson S (1996) Combining KOHONEN maps with ARIMA time series models to forecast traffic flow. *Transp Res C* 4:307–318
- Van Lint JWC, Van der Zijpp NJ (2003) Improving a travel time estimation algorithm by using dual loop detectors. *Transp Res Rec* 1855:41–48
- Van Lint JWC, Hoogendoorn P, Van Zuylen HJ (2002) Freeway travel time prediction with state-space neural networks-modeling state-space dynamics with recurrent neural networks. *Transp Res Rec* 1811:30–39
- Venkatanarayana R, Smith BL, Demetsky MJ (2005) Traffic pattern identification using wavelets transforms. In: 84th Transportation Research Board annual meeting, Washington, DC
- Vlahogianni EI, Golias JC, Karlaftis MG (2004) Short-term traffic forecasting: overview of objectives and methods. *Transp Rev* 24(5):533–557
- Vlahogianni EI, Karlaftis MG, Golias JC (2006) Statistical methods for detecting non-linearity and non-stationarity in univariate short-term time-series of traffic volume. *Transp Res C* 14(5):351–367
- Wahle J, Bazzan A, Klügl F, Schreckenberg M (2000) Anticipatory traffic forecast using multi-agent techniques. In: Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) *Traffic and granular flow '99*. Springer, Berlin Heidelberg, pp 87–92
- Wang Y, Papageorgiou M (2005) Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transp Res B* 39:141–167
- Webster FV (1958) Traffic signal settings. Road Research Laboratory technical paper no. 39
- Whitham G (1974) *Linear and nonlinear waves*. Wiley, New York
- Wiedemann R (1974) Simulation of traffic flow. Schriftenreihe des Instituts für Verkehrswesen der Universität Karlsruhe, Heft 8. (in German)
- Wild D (1997) Short-term forecasting based on a transformation and classification of traffic volume time series. *Int J Forecast* 13:63–72
- Williams BM (2001) Multivariate vehicular traffic flow prediction: an evaluation of ARIMAX modeling. *Transp Res Rec* 1776:194–200
- Williams BM, Hoel LA (2003) Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: theoretical basis and empirical results. *J Transp Eng* 129(6):664–672
- Williams JC, Mahmassani HS, Herman R (1987) Urban network flow models. *Transp Res Rec* 1112:78–88
- Xiao H, Sun H, Ran B, Oh Y (2003) Fuzzy-neural network traffic prediction with wavelet decomposition. *Transp Res Rec* 1836:16–20
- Yang F, Sun H, Tao Y, Ran B (2004a) Temporal difference learning with recurrent neural network in multi-step ahead freeway speed prediction. In: 83rd Transportation Research Board annual meeting, Washington, DC
- Yang F, Lin Z, Liu HX, Ran B (2004b) Online recursive algorithm for short-term traffic prediction. *Transp Res Rec* 1879:1–9
- Yasdi R (1999) Prediction of road traffic using a neural network. *Neural Comput Appl* 8:135–142. Springer
- Yin H, Wong SC, Xu J (2002) Urban traffic prediction using a fuzzy-neural approach. *Transp Res C* 10: 85–98
- Zhang HM (2000) Recursive prediction of traffic conditions with neural networks. *J Transp Eng* 126(6):472–481
- Zhang X, Rice J (2003) Short term travel time prediction. *Transp Res C* 11:187–210
- Zhang G, Patuwo E, Hu MY (1998) Forecasting with artificial neural networks: the state of the art. *Int J Forecast* 14:35–62
- Zwahlen HT, Russ A (2002) Evaluation of the accuracy of a real-time travel time prediction system in a freeway construction work zone. *Transp Res Rec* 1803:87–93

Books and Reviews

- Kalman R (1960) A new approach to linear filtering and prediction problems. *ASME Basic Eng J* 82(1):35–45
- Kants H, Schreiber T (2004) *Nonlinear time series analysis*. Cambridge University Press, Cambridge, UK

Physics of Mind and Car-Following Problem

Ihor Lubashevsky and Kaito Morimura
University of Aizu, Tsuruga, Ikki-machi, Aizu-
Wakamatsu, Fukushima, Japan

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Phenomenology of Human Actions and Physics
of Mind
Newtonian Approach to Car-Following
Three-Phase Theory and Car-Following
Car-Following: Driving Simulator Experiments
Car-Following: Dynamical Trap Model
Conclusion
Future Directions
Bibliography

Glossary

Dynamical trap is the fuzzy point matching the stationary point of a dynamical system described in terms of differential equations. The dynamical trap may be treated as a generalization of stationary point for systems governed by humans and corresponds to anomalously long residence time for systems entering the region of dynamical trap.

Intentionality is a property attributed to the mind and characterizing mental states as being “about something” with a functional reference to the intentional objects. Here *intentionality* is understood as the aspect of mental states characterizing them as “being goal-oriented” with some sequence of actions to be implemented for achieving the desired goal. In phenomenology intentionality is treated as a basic feature attributed to the mind and endowing *imaginary* future with causal power.

Intermittency of human actions is a generic term characterizing human actions that can be conceived of as an irregular sequence of alternate phases of active and passive behavior of a person in governing a certain system. This type of behavior appears when the dynamics of a system controlled by humans is stable outside the corresponding dynamical trap.

Main control parameter is a subjective phase variable whose variations are controlled completely by human volition actions and changing which humans can intentionally affect the other phase variables.

Phenomenology is the branch of philosophy of mind that studies conscious experience as experienced from the subjective or first-person point of view. In the framework of phenomenology, mental images representing observed objects in the mind are described in terms as they appear in the mind and how they are perceived or evaluated from the first-person point of view.

Physics of mind is a novel branch of physics studying mental phenomena based on the first principles of the mind and their mathematical formulations as well as the properties and dynamics of systems governed by humans. It turns to mechanisms of emotions, cognition, concepts, intuitions, imagination, etc., as well as takes into account human intentionality, goal-oriented behavior, volition, memory effects, and the bounded capacity of human cognition in the framework of phenomenology.

Space-time cloud is the image in the mind representing a point-like object taken at some instant of time. A space-time cloud has finite dimensions in space and time which are caused by the bounded capacity of discrimination between closely located spatial points and instants of time as well as conscious identification of such points and time moments. The *complex present* is the temporal component of the space-time cloud related to the present; the *fuzzy point* is its spatial component.

Two-component approach to describing systems governed by humans takes the central place in physics of mind. Within this approach, a system in issue is conceived of as a composition of two complementary components, *objective* and *subjective* ones, interacting with each other. The *objective component* comprises “material” constituent elements of a given system and obeys the law of classical physics. The *subjective component* represents the inner world of humans governing the system dynamics and obeys its own laws reflecting the basic features of the mind. As the main premise, physics of mind supposes that the quantities describing the states of objective component (*objective phase variables*) together with the quantities that can be introduced in the framework of phenomenology to describe subjective component (*subjective phase variables*) can provide a *closed* description of such a system.

Volition is the power or act of making a conscious choice or decision for intrinsic reasons. In phenomenology volition is treated as a basic feature attributed to the mind and endowing it with irreducibility (within phenomenology) to the physical reality.

Definition of the Subject and Its Importance

In developing a mathematical description of systems governed by humans, for example, ensembles of vehicles driven by humans on highways, we inevitably face two realities. One of them is the physical reality. If the positions of accelerator and brake pedals along with the rotation angle of steering wheel and the gear lever are given, the car motion is completely determined by physical laws, in a simplified form, by the laws of Newtonian mechanics. However this physical description is not complete. The position of control units is determined by *volitional* actions of car drivers, and there is no *direct* relationship between the spatial arrangement of vehicles and their velocities, on one side, and their representation in the human mind, on the other side.

It is the place where another reality – our inner world with its own regularities – comes on the scene. Actions of drivers depend on many factors, in particular, how the drivers perceive the current situation in the vicinity of their cars, their ability to anticipate the car arrangement in the near future, the individual goals the drivers intend to achieve in the future, the gained experience of driving, fatigue, the social norms of driving, etc. The given factors endow driver behavior with complex, nonlocal-in-time properties. Indeed, drivers in governing car motion respond to current stimuli, but how they respond is determined by these factors and changes in time according to psychological laws. In this case the personal history retained in memory and the expected future have causal power on the current actions. Leaving aside the philosophical long-term debates about the reducibility of the mental to the physical, we may confine our consideration to the approach dealing with physical states of car motion and the driver perception of these states as they appear in the mind. In this case the psychological laws governing the mental representation of car ensemble have to be regarded as irreducible to mechanical laws governing the car motion in the physical reality.

In other words, in modeling a system governed by humans, we need to deal with two realities, objective and subjective ones. The former represents the dynamics of the system affected by human actions in the objective space according to the laws of physics. The latter is formed by the human perception and cognitive evaluation of system states. The properties of the subjective reality reflect that of the objective reality but are not enslaved by them and exhibit their own dynamics. This approach is called phenomenology of human actions in philosophy of mind.

In the present entry, we employ the gist of phenomenological approach for constructing a model for the car-following, a paradigmatic example of intentional human actions in governing mechanical systems. In this way, we hope (i) to elucidate the basic features of subjective regularities and (ii) to find answers to some challenging problems in traffic flow physics.

Introduction

The present entry focuses the attention on the necessity of developing a new mathematical formalism for describing the dynamics of systems governed by humans. This formalism should explicitly allow for human memory, willingness to achieve some goals in future, emotions, social and ethic norms, etc., which are not under consideration within the paradigm of classical physics.

A promising approach to developing this formalism is to accept the gist of phenomenology of human actions developed in philosophy of mind. It opens a gate to the introduction of two complementary components interacting with each other, the objective and subjective ones. The objective component is governed by the physical laws; the subjective component representing the inner world of humans is governed by its own laws distinct from the laws of classical physics. The appearance of new objects of investigation – systems where human factor plays a crucial role – requires the development of a new branch of science which may be called physics of mind.

The concept of complex present in the subjective component – the human perception of the current moment of time – is the cornerstone of the given entry. To cope with complex present, the notion of space-time cloud and its special version – dynamical trap – will be introduced. The space-time cloud is the basic element representing the mental image of some event affected the bounded capacity of human cognition, delays in the mental processing of observed phenomena, the ability of unconscious short-term anticipation, etc. In other words, the space-time cloud is the mental image of a point-like event in the physical reality allowing for the space-time uncertainty of human perception and treated as some indivisible unit.

The problem of car-following is analyzed from this point of view. In particular, based on the conducted experiments on car-following using a car-driving simulator, a special dynamical trap model is proposed. It deals with a four-dimensional phase space including, first, the position of following car and its speed and acceleration which are the phase variables of the objective

component. Second, the last phase variable – the time derivative of car pedal position – is the main control parameter belonging to the subjective component and governed by volitional actions of drivers. The proposed model is demonstrated to be able to reproduce the characteristic features of human behavior, in particular, intermittency in human actions in controlling car motion.

Phenomenology of Human Actions and Physics of Mind

In constructing any mathematical model for human actions, we face up to several general questions which should be clarified before passing to particular investigations. In this section, we discuss two of them and elucidate the gist of approach to be used further in modeling driver behavior within the car-following scenario.

Basics of Mathematical Phenomenology

Dealing with individual behavior of humans, we meet two challenging problems standing in the way of developing mathematical models for such objects. One of them is *human individuality* raising doubts about the feasibility of describing human actions in mathematical terms. The other is *intentionality* of human behavior directed toward achieving some goals in future, including also the immediate future. This goal-oriented behavior of humans does not meet the paradigm of modern physics and its mathematical formalism.

Human individuality: We all are different, and sometimes this difference is so essential that we even do not understand one another. Does it imply that an individual “model” has to be created for each person? If so, the situation is hopeless because it is just impossible and, moreover, useless. One of the possible ways to tackle this problem is to confine our consideration to such systems whose behavior is not too sensitive to individual features changing from person to person. These objects which may be referred to as *cooperative social systems* are characterized by the following features (Lubashevsky 2017, cf. also Castellano et al. 2009; Oullier and Kelso 2009):

- Each system of this class is a large ensemble of humans joined by some common activity.
- It has no hierarchical, at least, fixed hierarchical structure.
- An individual can enter or leave this system freely.
- There is a certain mechanism coordinating the behavior of its members so that they have to obey some common rules.

For such a system at almost all time instants, only the “typical” actions of its members contribute substantially to system behavior. The human individuality is responsible for rare events admitting the interpretation in terms of mutually independent random events.

These features enable us to introduce the notion of *characteristic element* based on the concept of *local self-averaging* which holds for cooperative social systems (for details see Lubashevsky 2017). In such systems, the behavior of neighboring elements is strongly correlated, which allows us to deal with their large clusters in describing various cooperative phenomena rather than individual elements separately. Within these clusters, the system elements behave in similar way, and a characteristic element actually represents the properties attributed to the corresponding cluster as a whole. In other words,

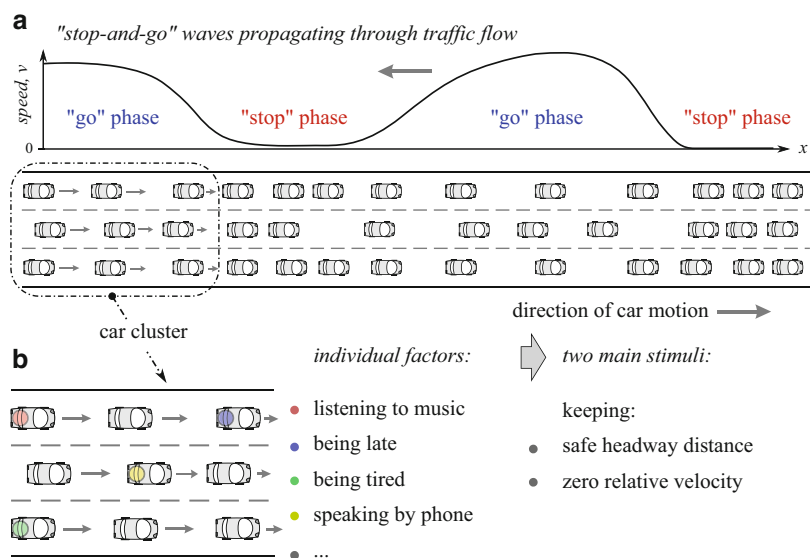
in describing cooperative social systems, the local self-averaging enables us to replace their real members by the corresponding characteristic elements. Because the number of *different* characteristic elements is rather small, the mathematical description of such objects becomes feasible.

It should be emphasized that there can be other mechanisms enabling the introduction of characteristic elements even for systems with weak interaction between their members or consisting of a few elements (Lubashevsky 2017). It opens a gate toward developing mathematical models for the individual behavior of real humans. In particular, the car-following to be considered below may be treated not only as a statistical representative of car interaction in traffic flow. Its mathematical model also admits the interpretation as a description of real actions of drivers. In this case, the model predictions may be verified by turning to the results of direct experiments on car-following.

To elucidate the basic features of cooperative social systems and the introduction of characteristic elements, let us turn to congested traffic flow on highways as a paradigmatic example. Figure 1a illustrates an ensemble of cars moving on a linear highway fragment when the car density becomes rather high and the “stop-and-go” waves arise. According to the modern state of the art in traffic flow physics (e.g., Kerner 2004, 2009a,

Physics of Mind and Car-Following Problem,

Fig. 1 “Stop-and-go” waves arising in congested traffic flow and the formation of car clusters with strong correlations in motion of their members



2017; Treiber and Kesting 2013), in such clusters cars interact with one another rather strongly such that their motion even on different lanes is not mutually independent. Each car has no room for maneuvering freely, and its driver has to synchronize the motion of his car with neighboring cars. This synchronization enables us to speak about car clusters as some structural units of traffic flow and to ascribe to them certain properties characterizing these clusters as a whole.

There are many factors as agitation, fatigue, vigor, skill of driving, etc. that affect the individual behavior of drivers (Fig. 1b). For example, if one is listening to a piece of quiet classical music, he can be relaxed and drive his car rather carefully without making risky maneuvers. If another driver is tired, his reaction is delayed, and attention to the motion of neighboring vehicles is weakened. When a third driver is late for some event important for him, he is nervous and can be ready for risky maneuvers.

In cases when the car density is rather low, vehicles practically do not hinder the motion of one another, and each person can choose any safe style of driving that is convenient for him individually. However, as the car density grows, the variety of driving styles acceptable under given conditions becomes more and more narrow, and thereby, all the drivers have to behave in a similar way. In this case, it becomes possible to confine our analysis to the consideration of a few stimuli governing the driver behavior and use a limited number of variables to quantify these stimuli. Either other factors like ones noted above are taken into account implicitly via the particular properties ascribed to these stimuli or their influence becomes of minor importance. In particular, dealing with traffic flow on linear fragments of highways provided overtaking maneuvers are rare events, we may consider only two main stimuli. They are (i) to keep a safe distance between the vehicles and (ii) to drive at the speed coinciding with the mean speed of the surrounding cars, mainly with the speed of the car ahead.

Other paradigmatic examples of cooperative social systems are crowd on public areas, dynamic social groups like sport fans, network of research collaboration, etc.

Thus, confining our consideration to cooperative social systems or, speaking more generally, to systems admitting the introduction of characteristic elements, we gain the possibility of constructing efficient mathematical models for describing human behavior and social phenomena. However overcoming the problem of human individuality in this way, we face up to another challenge.

Mathematical formalism for human intentionality: As far as modeling human behavior and social phenomena is concerned, whether new notions and formalism should be developed in addition to ones inherited from physics is up to now a challenging question. It is necessary to emphasize that the raised issue is not about the applicability of the basic physical notions like momentum, energy, and interaction potential to modeling social phenomena. It is about whether the general *formalism* of classical physics including differential equations with second-order time derivatives, the concept of system states, initial conditions, etc. can provide us with the mathematical vocabulary required for describing mental processes.

As an argument for this necessity, we note that in explaining human actions, we use a number of fundamental notions like long-term memory of past events evaluated in some way in the mind, free will, intentions, ethic norms, etc., which are just inapplicable to inanimate objects. So, in spite of the success achieved in modeling various human actions and social phenomena during the last decades, physics as a science about the inanimate world lacks experience of coping with humans especially in situations where these features play an essential role.

The basic premises underlying classical physics (actually Newtonian mechanics) provide us with a stronger argument for the necessity of new formalism which, moreover, has to meet another paradigm. As can be demonstrated (Lubashevsky 2017), classical physics is actually grounded on a general principle which may be referred to as the *principle of microlevel reducibility*. It comprises two theses:

- Only the present matters to physical objects because in the physical reality the past no longer exists and the future does not exist yet. In

philosophy this thesis is called presentism (for an introduction to the presentism and related concepts, see, e.g., Bourne (2006) and Hawley (2015)).

- Any complex system belonging to the realm of classical physics is nothing but an ensemble of structureless (point-like) particles interacting with one another. In addition, this interaction is supposed to obey the superposition principle or to admit a description in terms of linear fields locally interacting with point-like particles.

Quantum mechanics does not meet this principle, which however is not essential for our discussion because the physical component of systems in question can be described adequately in the framework of classical physics.

Here the term “structureless” may be understood in a broad sense; either these particles are really structureless or their structure does not change in time during analyzed processes and, so, can be treated as a fixed particle property. In any case, a mathematical formalism that is able to describe ensembles of classical particles within the quasi-stationary approximation of interaction fields deals with point-like objects (material points) whose properties are completely determined at a given instant of time. Dynamic properties of such material points are specified by the position in the phase space comprising their positions in the real space and their velocities. It should be noted that this determinism does not exclude the effect of dynamical chaos found in numerical simulation; the given effect is caused by some uncertainty in the position of a nonlinear system in its phase space.

If a system resides in some region S of its phase space for a relatively long time and a phenomenon under consideration is not sensitive to particular details of the system localization in S , the given region may be referred to as the current state S of this system. In the given case, it is possible to assume the analyzed phenomenon to be determined completely by the properties of the state S treated as a point in the space of possible states.

Whether the mental is reducible to the physical in various senses or, in particular, whether it is possible to *derive* models for human behavior

starting from the basic physical laws is a subject of long-term debates. Up to now there has been no answer found to this question arguing for or against it in an unambiguous, direct way. This situation is reflected also in the modern state of the art in physics of mind.

Broadly speaking, in physics of mind there can be singled out two quite distinct approaches to describing mental phenomena. One of them put forward by Perlovsky (2006, 2016) and Schoeller et al. (2018) accepts that the mind and brain refer to the same physical system at different levels of description. The mind includes conscious and partly unconscious processes, thought, perception, emotion, will, memory, etc. Specific neural mechanisms in the brain implementing these mind functions constitute the relationship between the mind and brain and may be used for constructing the mathematical description of mental phenomena turning to neural structure of the brain.

The other approach to be used below is focused on various aspects of mental phenomena that are accessible for human cognition and consciousness without referring to neural processes in the brain (Lubashevsky 2017). In this case, a plausible mathematical formalism is not necessarily to be confined to the paradigm of physics of the inanimate world and may operate with notions and concepts applicable only to the inner world of human beings. Such mathematical description of particular phenomena and regularities related to conscious has much in common with the approach called *phenomenology* in philosophy of mind. Phenomenology was launched in the first half of the twentieth century by Edmund Husserl and Martin Heidegger, et al. Broadly speaking, phenomenology studies structures of conscious experience as experienced from the *first-person point of view* (e.g., Smith 2013; Gallagher and Zahavi 2012, for introduction). To emphasize the crucial role of human self-awareness in developing such formalism, we will use the term *physics of the human mind* with respect to the given understanding of physics of mind. The former understanding of physics of mind does not exclude the consideration of various aspects of *animal mind* – the issue also discussed in modern literature (e.g., Andrews 2015).

In further mathematical constructions, the basic phenomenological concepts are employed; in particular, human individuals are regarded as the *basic entities* forming social systems. The term “basic entities” is used to emphasize that humans are treated as whole indivisible objects bearing all the properties ascribed to them.

Thus, the phenomenological approach grants us the freedom to go beyond the frameworks of classical physics in modeling the properties of human mind. Following this way, we may turn to a mathematical formalism admitting the goal-oriented behavior and allowing the past and future events to have causal power. As far as the car-following is concerned, the phenomenological approach, in particular, allows us to go beyond the formalism where the desired headway distance and the zero relative velocity specify some stationary point of a dynamical system.

Physics of the Human Mind

Within the phenomenological description two kinds of variables have to be introduced. First, it is the *objective* variables that (i) are directly accessible for the external observer and (ii) cannot be substantially changed by humans volitionally in no time or, speaking more strictly, on time scales of physiological delay in human response. The position of a car on a road and its velocity exemplify the objective variables. Second, it is the *subjective* variables characterizing human *intentions*. These variables can be inaccessible for the external observer; in this case, they just characterize the interval world of a given individual and may be referred to as hidden variables. The subjective variables can be accessible for the external observer, but in this case their values and changes are determined directly by individual's volition. The desired headway distance between a given car and the car ahead its driver wants to reach and the current position of gas or brake pedal illustrate subjective variables.

What regularities govern the dynamics of objective and subjective variables is actually the problem taking the central place in each phenomenological theory of human actions. To tackle this problem, Lubashevsky (2017) has proposed the concept of *two-component description* of

human actions which may be regarded as a particular implementation of the general phenomenological approach. Within this description, two *complementary* components – *objective* and *subjective* ones – are introduced. The objective component represents the physical (outer) world governed by the classic laws of physics. The subjective component represents our inner world and is assumed to be governed by its own laws. Naturally, the objective and subjective components affect each other. The subjective variables accessible for the external observer also belong to a “bridge” between the two components specifying their interaction.

Within the given approach, the laws of the subjective component are irreducible to the laws of objective component. This irreducibility becomes evident just pointing to the fact that the principle of microlevel reducibility does not hold in describing human actions within frameworks where humans are treated as the *basic entities*. Indeed, the behavior of each individual is directly affected by the goal he pursues, his experience related to the current situation, its evaluation and recognition, etc. So the subjective component has a complex temporal structure involving:

- The subjective present, i.e., the human perception of the current state of objective component
- The past retained in memory and the body experience
- The imaginary future with intentional actions that should be implemented to get the desired goals

These features are not directly accessible for the external observer and, moreover, cannot be regarded as localized at some instant of time. Even the present perceived by humans as some instant of time has a finite temporal dimension in the reality, which is reflected in the notion of *specious present*, the finite time interval whose points are perceived by humans as simultaneous. The term *specious present* was coined by E. Robert Kelly (alias “E.R. Clay”) (1882) and further developed by philosopher and psychologist William James (1890) (see, for details, Andersen and Grush 2009). Moreover, as explained below, the *specious*

present possesses some internal structure, and the use of the *complex present* is preferable to emphasize this fact (Lubashevsky 2017).

Figure 2 illustrates the gist of two-component description. The objective component is governed by the classical laws of physics and the present instant of its time is a pointlike object. The subjective component obeys its own laws irreducible to the paradigm of classical physics and possesses a complex temporal structure. Therefore the governing laws of subjective components, i.e., the corresponding regularities of mental phenomenon, have to be described in the framework of individual formalism and studied by methods developed specially for these purposes.

The two-component approach supposes that human behavior can be described in a *closed way* using (i) the physical laws of objective components, (ii) the individual laws of subjective components, and (iii) the laws of the interaction between the two components.

Thus. physics of mind is a new branch of physics that studies properties and dynamics of systems governed by humans and takes into account intentions, goal-oriented actions, memory effects, features of human perception, and mental phenomena within the phenomenological approach.

Concept of Space-Time Clouds

The temporal structure of complex present and its corresponding spatial structure are reflected in the

concept of *space-time clouds* treated as basic entities of the subjective component (Lubashevsky 2017).

A *space-time cloud* is the mental image of a point-like physical object (material point) after its imaginary embedding into the physical space-time continuum. In other words, in our mental estimation of space-time arrangement of material points, we deal with the corresponding space-time clouds in the mind. Within the framework of two-component description, the space-time cloud is the image of the material point – the element of objective component – in the subjective component. Controlling, for example, the dynamics of some physical system composed of material points, we perceive this system as an ensemble of space-time clouds. So our actions are determined by these clouds but affect the physical objects. Figure 3 illustrates these constructions.

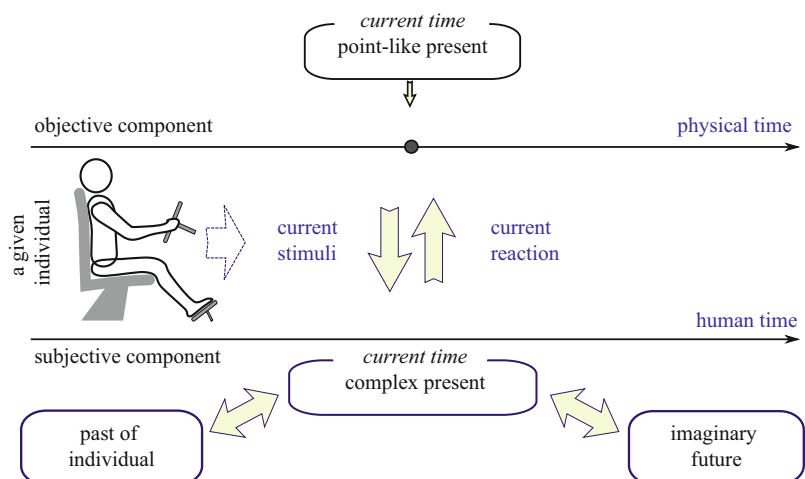
In contrast to the material point, the space-time cloud is characterized by

- Finite extension in space
- Finite duration
- Fuzzy boundaries in space and time
- Internal dynamics related to human cognition of a real point-like object under particular conditions

which is illustrated in Fig. 4. Time uncertainty caused by the finite duration of complex present can be evaluated as a quantity bounded from below by 50 ms and from above by 1 s depending

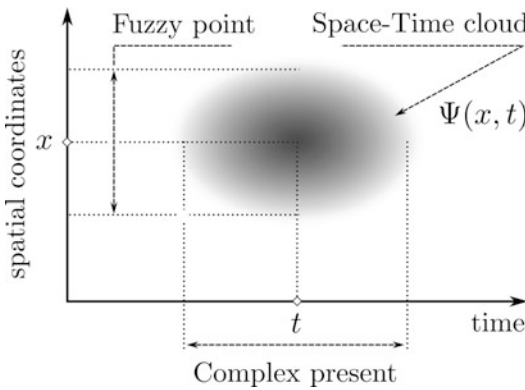
Physics of Mind and Car-Following Problem,

Fig. 2 Diagram representing the details of the interaction between the objective and subjective components of a given individual and the interaction between the different temporal parts of the subjective component



Physics of Mind and Car-Following Problem,

Fig. 3 Illustration of the relationship between the elements (material points) of the objective component, their images in the mind, and the corresponding space-time clouds in the subjective component being the results of embedding the mental images into the mental image of the space-time continuum



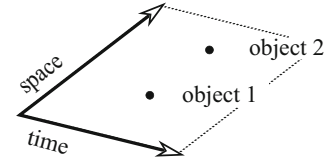
Physics of Mind and Car-Following Problem, Fig. 4 Illustration of the space-time cloud – the notion combining in itself the notions of complex present and fuzzy point characterizing the human perception of point-like objects

on particular conditions (see Elliott and Giersch 2016, for a review). This interval also represents the typical values of the human reaction delay time found in experiments on human response to initially known stimuli or involving recognition tasks too (e.g., Proctor and Vu 2003).

The *spatial uncertainty* of space-time cloud is reflected in the notion of *fuzzy point* (Fig. 4). The fuzzy point and the complex present are the constituent parts of space-time cloud, and some of their properties can be attributed to the space-time cloud as a whole, which endows it with integrity. This integrity stems from the fact that events neighboring in space and time may be *perceived* by humans as *coincident*. Moreover, such objective component events may be *treated*

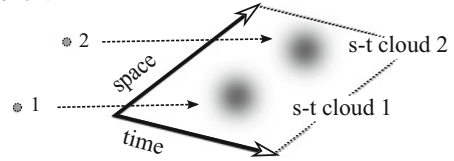
objective component

object's arrangement in physical space



subjective component

object's images in mind



as coincident because their contributions to analyzed phenomena are practically equivalent and there is no need to consider them separately.

In what follows let us confine our discussion to a special case related to the dynamics of some system governed by human actions aimed at keeping the system near a desired (equilibrium) state. In particular, car-driving in a traffic flow exemplifies this situation.

Classical approach to self-organization: Within the formalism of dynamical systems widely employed in modern physics of self-organization phenomena (see, e.g., Haken 2009a, b), the motion of a nonlinear system is described by differential equations:

$$\frac{dq_i}{dt} = \mathcal{F}_i(\mathbf{q}) + \text{random source}, \quad (1)$$

where $\mathbf{q} = \{q_i\}$ is the point of system phase space, the given functions $\mathcal{F}_i(\mathbf{q})$ called *generalized forces* or just *forces* for short are assumed here not to depend on time t , and the right-hand side of Eq. (1) may contain the noise component allowing for possible random forces affecting the system motion. The notion of *stationary point* $\mathbf{q}^{\text{st}}$ defined via the equality

$$\mathcal{F}_i(\mathbf{q} = \mathbf{q}^{\text{st}}) = 0 \quad \text{for all } i \quad (2)$$

takes one of the central places in the theory of self-organization. When the stationary point loses its stability, the system goes way from it, leading to the formation of various spatiotemporal patterns.

If the stationary point is stable, the system will reside in its small neighborhood until something happens, and it loses the stability or a rare strong noise fluctuation causes the system to go out the stability region of this point.

The situation changes drastically when the system motion toward the point $\mathbf{q}^{\text{st}}$ as a desired state of the given system is controlled by human actions. For example, turning to Fig. 3, let us assume that a human tries to drive object 1 toward object 2 to place it in some imaginary neighborhood of object 2. For him the system gets its desired state $\mathbf{q}^{\text{st}}$ when clouds 1 and 2 overlap substantially and further correction of system state is not necessary. These constructions prompt us to generalize the notion of stationary point $\mathbf{q}^{\text{st}}$ to allow for the bounded capacity of human cognition in the following way.

Dynamical trap as stationary point generalization: In order to describe the effects of spatio-temporal uncertainty in human evaluation of the system proximity to the desired state $\mathbf{q}^{\text{st}}$, the notion of *dynamical trap* is required. The *dynamical trap* is the image $\mathbb{Q}_{\text{tr}}$ of the stationary point $\mathbf{q}^{\text{st}}$ in mind, i.e., the point $\mathbf{q}^{\text{st}}$ treated as the fuzzy point.

The fuzziness of $\mathbb{Q}_{\text{tr}}$ endows the human decision about getting the desired state with uncertainty and, thereby, the system dynamics with probabilistic properties. When the system enters the dynamic trap region $\mathbb{Q}_{\text{tr}}$, the human actions on driving the system are halted, and its dynamics can be stagnated for a long time. Thereby the dynamical trap $\mathbb{Q}_{\text{tr}}$ inherits the main property of stationary point – the possibility of the system residing in it, at least, for anomalously long time interval.

By way of example, let us consider the car-following to elucidate the introduction of dynamical traps. In following a lead car, each driver has continuously (i) to control the speed difference between his car and the car ahead and (ii) to keep a safe headway distance. However, the accuracy of the two processes is remarkably different. The driver of the following car easily perceives the velocity difference about 2–5 km/h (Reiter 1994; Brackstone and McDonald 1999), which is about 5% for the speed of 60 km/h. As far as

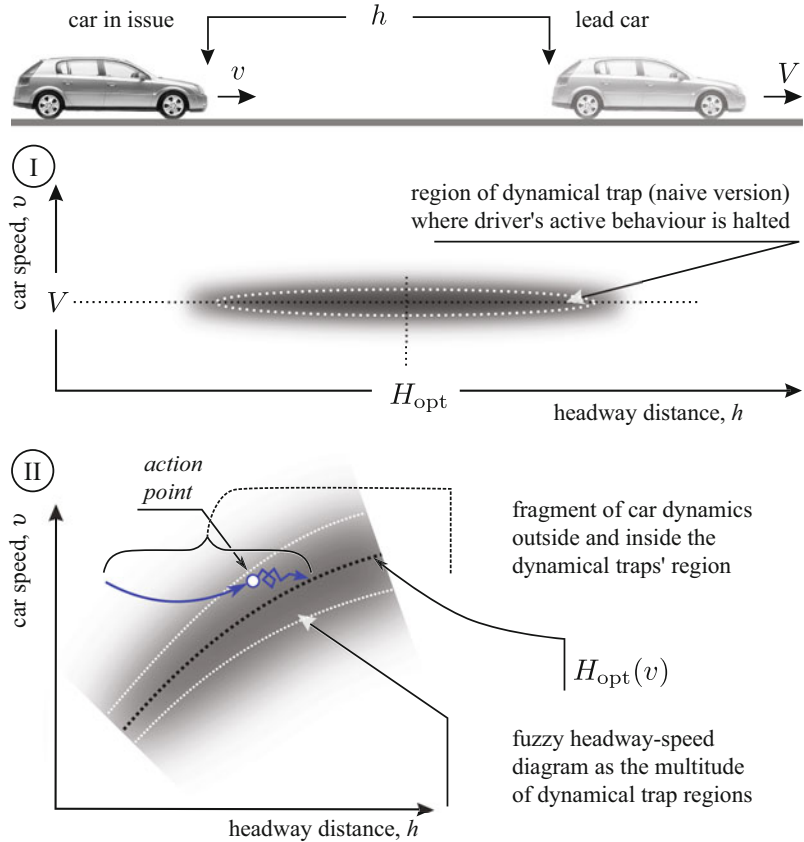
the accuracy of headway distances is concerned, the available experimental data (see, e.g., Treiber and Kesting 2013) show that 50% variations of headway in traffic flow on highways for a fixed speed are typical. So a certain, rather wide, multitude of headway distances is acceptable in car-following. As a result, when the following car gets the neighborhood of the lead car matching this headway distance multitude, the driver halts the further correction of the headway for a long time. So in describing such driver actions, we can single out some region – the region of dynamical trap in its naïve interpretation (Fig. 5I) – inside which the driver control over the car position is halted (passive phase). When the necessity of correcting the car position becomes evident, the driver changes the driving strategy and passes to the active phase. The concept of dynamical traps based on the available empirical data for congested traffic flow was also discussed by Lubashevsky et al. (2002b).

The given concept of dynamical trap may be also represented in the form more suitable for describing traffic flow in terms of fundamental diagram, in particular, the headway-speed diagram (Fig. 5II). This representation may be referred to as the *fuzzy headway-speed diagram*.

Let us consider stationary traffic flow on linear highway fragments which consists of one type of vehicles, e.g., passenger cars, and whose state is characterized by weak variations in car speed over ensemble. Synchronized traffic to be discussed below exemplifies such states of car ensemble. Then let $H_{\text{opt}}(v)$ denote the mean headway distance in such traffic flow depending on the mean car speed v . Here $H_{\text{opt}}(v)$ is calculated by averaging over time and car ensemble. Within Newtonian approach to modeling the car-following to be discussed below too, the quantity $H_{\text{opt}}(v)$ is often referred to as the optimal value of headway distance and treated as the stationary point of the corresponding governing equations for a given speed v of the lead car. On the *headway-speed* plane, $\mathbb{R}_{h,v}$, the locus of points $\{h = H_{\text{opt}}(v), v\}$ forms a certain curve usually called the headway-speed diagram for a given type traffic flow (Fig. 5II).

Physics of Mind and Car-Following Problem,

Fig. 5 Illustration of the dynamical trap describing the car-following problem when the driver perception of the motion of the car ahead is characterized by the bounded capacity of evaluating the headway distance



However, even if the distance $H_{opt}(v)$ is optimal in some way, a driver cannot identify precisely the current headway with the value of $H_{opt}(v)$; he can do this only approximately within accuracy of order of 50%, which is reflected in the concept of dynamical trap. The multitude of dynamical trap regions centered at the curve $h = H_{opt}(v)$ for possible values of v make up a two-dimensional domain $\mathbb{Q}_{hv}$ with blurred boundaries called the *fuzzy headway-speed diagram*.

The fuzzy headway-speed diagram inherits its properties from dynamical traps. Namely, when, during the correction of car position in traffic flow, the current state of car motion – its position $\{hv\}$ on the plane $\mathbb{R}_{hv}$ – gets the domain $\mathbb{Q}_{hv}$, the driver has no reason to correct further the headway distance. In this case, he needs only to keep up the car speed near the speed of lead car. The further dynamics of the car motion state is affected by uncontrollable random factors. Therefore if a given fragment of such traffic flow matches an

internal point of the domain $\mathbb{Q}_{hv}$, it will subsist for a long time. In other words, any point of the domain $\mathbb{Q}_{hv}$ represents a certain long-lived state of this type traffic flow.

The boundary fuzziness of $\mathbb{Q}_{hv}$ reflects the probabilistic nature of transitions between the driver intentional behavior aimed at decreasing or increasing the headway distance and the temporary halt in the headway control (Fig. 5II). These transitions – action points in Todosiev's terminology (Todosiev 1963; Todosiev and Barbosa 1963/64) – must be noise-induced phenomena (Zgonnikov et al. 2014; Zgonnikov and Lubashevsky 2015, see also Lubashevsky 2017) rather than governed by the threshold mechanism. Below we mainly will deal with a continuous representation of the dynamical trap effect (Lubashevsky 2012, 2016).

A more sophisticated description of the dynamical trap effect is based on the two-component approach. Let us outline it to show its general structure. A system governed by

human action is described by two types of variables, objective and subjective ones, $\mathbf{q} = \{q_i\}$ and $\mathbf{u} = \{u_k\}$, respectively, which make up its *extended phase space* $\mathbb{R}_{qu}$. The dynamics of the q -variables is assumed to be governed by some forces $\mathcal{F}_i(\mathbf{q}, \mathbf{u}, t)$ depending on both the q - and u -variables and, maybe, time t :

$$\frac{dq_i}{dt} = \mathcal{F}_i(\mathbf{q}, \mathbf{u}, t). \quad (3a)$$

In the general form, the dynamics of the u -variables is represented as

$$\frac{du_k}{dt} = \begin{cases} \mathcal{R}_k(\mathbf{q}, \mathbf{u}, t) & (\text{activephase}) \\ 0 & (\text{passivephase}) \end{cases} + \text{noice}, \quad (3b)$$

where the functions $\{\mathcal{R}_k(\mathbf{q}, \mathbf{u})\}$ are specified by human actions and the noise-related term allows for possible uncontrolled factors. The probabilistic transitions

$$\text{active phase} \Leftrightarrow \text{passive phase for } \{\mathbf{q}, \mathbf{u}\} \in \mathbb{Q}_{tr} \quad (3c)$$

are determined by the human decision-making when the system enters or leaves the dynamical trap region $\mathbb{Q}_{tr}$ being a certain fuzzy point of the extended phase space $\mathbb{R}_{qu}$.

It is necessary to note that there have been proposed two different notions of dynamics traps. One of them proposed by Zaslavsky (1995, 2002, see also a book by Zaslavsky, 2005) appears in the theory of Hamiltonian systems with complex dynamics and describes non-linear physical phenomena.

The other type of dynamical traps proposed by Lubashevsky et al. (1998, 2002b, 2003a) describes human actions and reflects their basic features. The dynamical traps of this kind can give rise to a novel type of self-organization phenomena which are not caused by the instability of stationary points (Lubashevsky 2016, for a review).

Intermittency in Human Behavior

The effect of dynamical traps in driving a system near its equilibrium has to endow human actions

with some *intermittency*. Here intermittency means irregular transitions between the passive and active phases of behavior. By way of example, let us note two particular phenomena well studied nowadays where this intermittency is recognized as their essential property. The model for car-following to be presented below also turns to the concept of intermittency as a basic feature of driver behavior.

Human intermittent control is a novel concept of describing human actions in governing various systems (for a review, see, e.g., Gawthrop et al. 2011; Loram et al. 2011; Balasubramaniam 2013; Milton 2013; Asai et al. 2013). Such human actions comprise active and passive phases of behavior, with the active phase being of the open-loop control type. As a widely used approximation, the transitions between the active and passive phases are assumed to be determined by events when the deviation of governed system from its equilibrium state crosses some threshold. Besides, human response to changes in the system state can be delayed for some fixed time-shift. This event-driven model may be treated as a certain approximation of space-time clouds with strict boundaries.

Another approach to describing the transitions between active and passive phases of human control has been proposed by Bottaro et al. (2008), Asai et al. (2009, 2013), Suzuki et al. (2012), and Yoshikawa et al. (2016). It supposes that in the corresponding phase space, there is some manifold with two branches, stable and unstable ones. The system motion along the stable branch leads to the desired equilibrium on its own, i.e., without any human interference in the system dynamics. Therefore human control becomes active only when the system deviates from the stable branch of this manifold and is aimed at returning the system to this branch rather than driving it to the desired equilibrium directly. It should be noted that this concept of human control meets the general properties of goal-oriented actions combined with the presentation of active phase as open-loop control (e.g., Zgonnikov et al. 2014; Lubashevsky 2017). For modeling human goal-oriented actions, the formalism dealing with higher-order time derivatives is required. The use of this

formalism leads to the introduction of an extended phase space, the stable and unstable manifolds of system motion near the corresponding stationary point, and, finally, to the turn to the concept of dynamical traps; with respect to the car-following problem, this approach is implemented by Lubashevsky et al. (2002a, b, 2003b, c).

Recently a novel mechanism of control activation has been proposed by Zgonnikov et al. (2014). It suggests that control activation stems from stochastic interplay between the need to keep the controlled system near the desired state, on the one hand, and the tendency to postpone interrupting the system dynamics, on the other hand. This noise-induced control activation regards the physical and mental factors in such human actions as equipollent components with individual dynamics and own laws.

The fact that active phase corresponds to open-loop control allows to turn to a more general interpretation of intermittent control. Broadly speaking, during the passive as well as active phases, a person does not change the current actions. In the former case, he does not affect the system motion at all; in the latter case, he does not affect the stream of planned actions, i.e., does not change the strategy chosen previously until it is not implemented completely. In this sense, both the phases may be treated as passive phases different in properties.

The active behavior of the person is reflected at the moment when he makes decision about changing the current state of system motion. In this way, we get the notion of *action points* introduced into the theory of traffic flow by Todosiev (1963) and Todosiev and Barbosa (1963/1964). For the modern state of the art in the theory of action points in driver behavior, a reader may be addressed to Wagner (2006, 2012), Hoogendoorn et al. (2011), Bifulco et al. (2013), and Pariota and Bifulco (2015).

Motor behavior: The intentionality also takes the central place in *motor behavior* – body movements aimed at achieving certain goals – (Latash 2012a, 2015; Huys et al. 2014, for an introduction).

The approaches proposed nowadays to modeling motor behavior can be approximately divided into two classes, *computational* and *physical* models. The *computational approach* is closely related to the concepts of cognitive psychology

treating the human mind as an information-processing device (see, e.g., Schmidt 1975, 2003; Schmidt and Lee 2011, for details). The *physical approach* to describing human motor behavior is based on the models combining the basic ideas developed in the biomechanics of neuromuscular system and the theory of self-organization phenomena (e.g., Latash 2008).

The physical approach inevitably faces up to the challenge of explaining goal-directed actions in physical terms because anticipation and evaluation of future events do not belong to the realm of physics (e.g., Clark 2013). There have proposed several ways to overcome this problem.

A general framework was put forward by Bernstein (1935) based on the notion of *engrams* as precursors of movements stored in memory. Engrams reflect topology of planned movements rather than their metrics. When the initial part of an engram has been implemented in a particular case, the implementation of the other parts will be also initiated in the appropriate way in advance of the future events. Nowadays Bernstein's ideas are reflected in the *equilibrium point theory* and the concept of *anticipatory motor adjustments* developed by Latash et al. (for details see Latash (2015) and references therein).

Another way the concept of *active inference framework* turns to the ideas of nonequilibrium thermodynamics. In its frameworks, the description of human actions is reduced to the minimization of a certain free energy. Thereby the goal-oriented decision-making is replaced by movement trajectories realized as random events (Friston et al. 2015; Pezzulo et al. 2015, for a review).

One more way is related to the formalism of differential equations. To overcome the problem of intentional actions and the high redundancy of body movements, the concept of *uncontrolled manifold* is accepted as the pivot point (Scholz and Schöner 1999; Latash et al. 2007; Martin et al. 2009, for details and available evidence, see overviews by Latash (2012b); Huys et al. (2014)). This concept assumes body movements can be represented as a system motion on a certain manifold embedded into a multidimensional space. A deviation from the current implementation of body movement caused, e.g., by a perturbation

within this manifold does not require correction and, thus, is not controlled. Only the system deviations from this manifold have to be corrected and are mainly controlled by the neuromuscular network.

Newtonian Approach to Car-Following

Ensemble of vehicles moving on a highway is a characteristic example of wide class of many-element systems often referred to as self-driven particles. A self-driven particle is a generic term standing for a living being in its motion for some intrinsic reasons. Pedestrians on crowded areas, flocks of birds, and schools of fish are other typical examples of this class. The self-organization of various spatiotemporal patterns in such systems has attracted much attention during the last decades. The intrinsic stimuli cause such entities to move, whereas the interaction with their neighbors induces synchronization of their motion and gives rise to self-organization phenomena in their ensembles.

Following the lead car is one of the main microlevel “bricks” making up the physics of traffic flow. So the understanding of its basic properties is necessary for developing sophisticated models for complex self-organization phenomena in traffic flow, including ones determining the limit capacity of highways as well as the street-road networks of big cities (see, e.g., Kerner 2004, 2009a, 2017).

A formalism rooted in the fundamentals of Newtonian mechanics and widely used in studying particular phenomena in such systems is outlined in the present section.

At the microlevel dealing with individual self-driven particles, their dynamics is described by the Newtonian-type differential equation:

$$\frac{d\mathbf{v}_i}{dt} = \mathbf{F}_i(\{\mathbf{X}_j, \mathbf{V}_j\}_j), \quad (4)$$

where $\mathbf{X}_i$ and $\mathbf{V}_i$ are the position and velocity of a given particle i . Now, following Lewin (1951), Helbing (1991, 1994), and Helbing and Molnár (1995), the functions $\mathbf{F}_i(\dots)$ depending on the position and velocities of all the particles are often called

social forces. This term was introduced to emphasize that such forces can be of social nature or stem from the rules of behavior the group members follow volitionally for their own reasons. Model (4) itself is also often referred to as the *social force model*. The term “social force model” is usually associated with modeling the dynamics of pedestrians, birds’ flocks, etc. rather than vehicle traffic. Nevertheless, here we use this term with respect to car-following also to emphasize the close relationship between the models to be noted in the given section and the paradigm of Newtonian mechanics.

In describing individual dynamics of a car moving in traffic flow, e.g., on highways, usually two factors are assumed to govern the driver behavior. The first one is keeping the velocity v_i of a given car i equal to the velocities $\{v_j\}$ of neighboring cars. The second one is maintaining the safe headway distance $|x_j - x_i|$. Both the factors cumulatively determine the car acceleration.

Turning to equations like Eq. (4), the dynamics of traffic flow on single-lane roads or congested traffic on highways where overtaking events are rare is typically described as

$$\left. \frac{dv_i}{dt} \right|_t = F_i(\{x_j, v_j\}_j)_{t-\tau}. \quad (5)$$

These models can take into account some time delay τ in driver response. Here the position x_i and velocity v_i of car i are one-dimensional variables, and the car arrangement can be characterized by a fixed order of cars

$$\dots x_{i-1} < x_i < x_{i+1} < \dots$$

along the road if the contribution of overtaking in car dynamics is ignorable. The approximation where the driver of car i reacts only to the behavior of car $i + 1$ located in front of him is called the *car-following model*. In this case, Equation (5) takes the form

$$\left. \frac{dv_i}{dt} \right|_t = F(\{x_{i+1} - x_i, v_{i+1}, v_i\}_j)_{t-\tau}. \quad (6)$$

where the forces acting on all the cars are often assumed to have the same functional form.

Models similar to (5) and (6) suppose the position of cars and their velocities to determine completely the car accelerations. In this sense, the given models share the basic premises of Newtonian mechanics in describing the dynamics of traffic flow; however, strictly speaking, they do not meet the paradigm of Newtonian mechanics when the delay of driver reaction plays an essential role.

Early models of this type were proposed by Reuschel (1950a, b) and Pipes (1953) treating the velocity difference $v_{i+1} - v_i$ as the main stimulus to driver actions. Later such models received the name *following-the-leader models* within which, in particular, the sensitivity factor can depend in a rather sophisticated way on the car velocity v_i and the headway distance $h_i = x_{i+1} - x_i$. For example, (Gazis et al. 1959, 1961; Edie 1961):

$$\left. \frac{dv_i}{dt} \right|_t = kv_i^m(t) \frac{v_{i+1} - v_i}{(x_{i+1} - x_i)^l} \Big|_{t-\tau}, \quad (7)$$

where the constant k and the exponents m and l are model parameters.

The development of models focusing their attention on the safe headway distance called the *optimal velocity models* was initiated by Newell (1961) and Bando et al. (1995, 1998). In particular, according to Bando et al. (1998)

$$\left. \frac{dv_i}{dt} \right|_t = k[\vartheta(x_{i+1} - x_i) - v_i] \Big|_{t-\tau}, \quad (8)$$

where $v(h)$ is the optimal velocity, i.e., the mean velocity accepted by drivers for a given headway distance h .

There have been proposed a number of models taking into account both the stimuli and characteristic features of human behavior, in particular, the *generalized forced model* (Helbing and Tilch 1998), the *full velocity difference model* (Jiang et al. 2001), and the *intelligent driver model* (Treiber et al. 2000). For example, the intelligent driver model reads

$$\begin{aligned} \frac{dv_i}{dt} = a & \\ & \cdot \left[1 - \left(\frac{v_n}{v_0} \right)^\delta + \left(\frac{s^*(v_i, v_{i+1} - v_i)}{x_{i+1} - x_i} \right)^2 \right], \end{aligned} \quad (9)$$

where the function

$$s^*(v_i, v_{i+1} - v_i) = s_0 + v_i T + \max \left[0, \frac{v_i(v_{i+1} - v_i)}{2\sqrt{ab}} \right]$$

and a , v_0 , δ , S_0 , T , and b are model parameters.

There are also models allowing for additional factors in the car-following. For example, the models by Gaididei et al. (2013) and Carter et al. (2014) take into account the individual dynamics of the headway distance perceived by drivers as optimal. The general concept of slow dynamics manifold put forward by Marschler et al. (2014, 2015) enables to analyze traffic flow properties without specifying a microlevel governing equations. Within the generalized full velocity difference models (Zhao and Gao 2005; Yu et al. 2013) governing equations similar to Eq. (6) contain also the acceleration of the lead car a_{i+1} or even the acceleration $a_i(t - \tau)$ of the given car but taken with some time-shift. The introduction of acceleration with time delay formally endows the corresponding models with non-Newtonian properties. Indeed in this case using the Taylor series we directly obtain terms containing higher order time derivatives. This, however, modifies mainly the conditions of the traffic flow instability rather than gives rise to novel phenomena (Nagatani and Nakanishi 1998).

There are also models often referred to as *cellular automata* that deal with individual vehicles whose motion is described in discrete time and space. The Nagel-Schreckenberg and Krauss models (Nagel and Schreckenberg 1992; Krauss et al. 1996) are their paradigmatic examples (for a review of this approach, see, e.g., Chowdhury et al. (2000), Mahnke et al. (2005), Maerivoet and Moor (2005), Treiber and Kesting (2013), and Li et al. (2016)).

Three-Phase Theory and Car-Following

The *three-phase traffic theory* developed by Kerner during the last two decades (for details see Kerner 2004, 2009a, 2017) has brought into physics of traffic flow the basic concepts of phenomenological theory of phase transitions in physical media. In particular, the notions of metastable states, critical nuclei, and spontaneous and induced breakdown are employed to describe complex emergent phenomena observed in traffic flow. The gist of three-phase traffic theory is that, first, there are three distinct phases of traffic flow (e.g., Kerner 2004, 2009a, b, 2017):

- *Free flow (F)*, the state of traffic flow when the car density is rather small. In this case, the car maneuvers required for overtaking are simple in implementation, and each person can freely drive his car with a speed meeting the weather conditions, speed limitations, and his own preferences.
- *Wide moving jam traffic phase (J)*, a wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks.
- *Synchronized flow traffic phase (S)*, the downstream front of synchronized flow does not exhibit the abovementioned characteristic feature of wide moving jams; specifically, the latter front is often fixed at a bottleneck. (The given definitions of wide moving jam and synchronized traffic are taken from Kerner 2009b, p. 9303):

The traffic phases “synchronized flow” and “wide moving jam” in congested traffic can also be distinguished through the *microscopic criterion* for traffic phases in congested traffic of Kerner et al. (2006).

Second, in the three-phase theory, traffic breakdown is a phase transition from free flow to synchronized flow ($F \rightarrow S$) which is governed by the nucleation mechanism endowing it with random properties. This is the fundamental feature of the three-phase theory. Wide moving jams can emerge spontaneously in synchronized flow via

the transition $S \rightarrow J$ which is also governed by the nucleation mechanism. Thus, in the three-phase theory, wide moving jams emerge spontaneously due to a sequence of two-phase transitions:

$$F \rightarrow S \rightarrow J. \quad (10)$$

The nucleation mechanism implies that a new phase arises when inside the current phase a cluster of new phase with size exceeding a certain critical value – a critical nucleus – appears due to random fluctuations. Then the further irreversible growth of critical nucleus causes the phase transition. In particular, exactly the former transition, $F \rightarrow S$, exhibiting substantially probabilistic behavior, determines the highway capacity (e.g., Kerner 2017, for a detailed discussion).

The synchronized traffic exhibits a number of anomalous properties, e.g., it matches a wide two-dimensional region on the plane *traffic flow rate – the car density* where each point represents some long-lived state of synchronized traffic (Kerner and Rehborn 1997). Treiber and Helbing (1999) coined a generic term – *widely scattered states* – for this region which allows for various plausible mechanisms such as difference in driver characteristics and vehicle parameters as well as speed and density disturbances always present in real traffic flow. However, it should be emphasized that this two-dimensional multitude of long-lived states is mainly caused by the fundamental properties of human behavior in car-driving rather than external effects like traffic flow heterogeneity. In particular, nowadays only the three-phase theory admitting the existence of volumetric multitude of long-lived states seems to be able to describe the observed probabilistic properties of traffic breakdown turning to the physics of phase transitions (10) (Kerner 2016; Kerner 2013).

Kerner (1998, 1999, 2004, 2009a, b) put forward the following hypotheses belonging to the fundamentals of the three-phase theory:

- Steady states of synchronized flow cover a two-dimensional (2D) region in the flow-density plane (and, correspondingly, in the headway-speed plane).

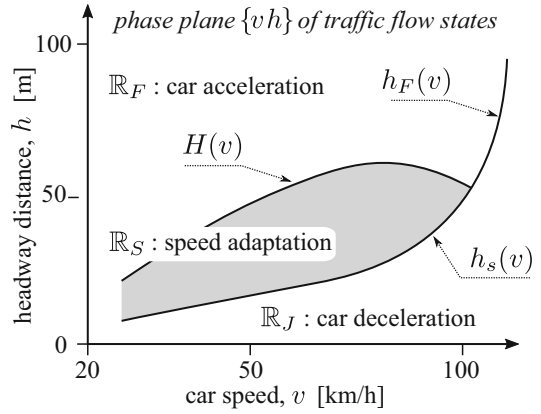
- At relatively small space gaps, which occur in synchronized flow, a driver can recognize whether the space gap to the preceding vehicle in car-following increases or decreases over time.
- At a given speed in synchronized flow, the driver can make an arbitrary choice in the space gap to the preceding vehicle within a finite range of space gaps associated with the 2D region of steady states of synchronized flow.
- [The] driver accepts different space gaps at different times and does not control a fixed space gap to the preceding vehicle (Kerner 2009a, pp. 46, 47, text in parentheses is added by us).

These hypotheses prompted Kerner et al. to propose a number of car-following models including, in particular, the *deterministic acceleration time delay three-phase traffic flow model* (ATD model) (Kerner and Klenov 2006), the *stochastic three-phase traffic flow model* (Kerner and Klenov 2002, 2003), and the *cellular automata three-phase traffic flow model* (Kerner et al. 2002) (for a review of this line, see, e.g., Kerner (2009a, 2017)). By way of example, let us confine our discussion to the ATM model; Fig. 6 illustrates its key point.

According to the three-phase theory, within some velocity interval (v_l, v_u) on, for example, the $\{vh\}$ -plane (plane *car speed-headway distance*), there can be singled out three regions $\mathbb{R}_F$, $\mathbb{R}_S$, and $\mathbb{R}_J$ (Fig. 6). The region $\mathbb{R}_F$ bounded from below by the curve $H(v)$ represents the situation when the headway distance h between the lead car moving at speed v_l and the car-following it exceeds some value $H(v)$. In this case, the driver of the following car tries to decrease this distance driving his car, on the average, with the acceleration:

$$a_{\text{opt}}(h, v, v_\ell) = A[V(h) - v] + K(v, v_\ell)[v_\ell - v],$$

where v is the current speed of the following car and the function $V(h)$ is the inverse of $\{h_F(v)\}$.



Physics of Mind and Car-Following Problem, Fig. 6 Characteristic regions of the phase plane *car speed-headway distance* that match different types of driver behavior, based on Kerner and Klenov (2006, Fig. 1)

However, for synchronized flow $h_s(v)$ is not some optimal headway taken by the majority of drivers in motion with speed v . Within the three-phase theory of traffic flow, drivers are assumed to consider any headway distance in the interval $h_s(v) < h < H(v)$ (region $\mathbb{R}_S$) acceptable. So after getting the region $\mathbb{R}_S$, they have no reason to correct the headway, and their actions are mainly aimed at keeping their speed v equal to the speed v_l of the car ahead. This speed adaptation is specified by the expression

$$a_{\text{opt}}(h, v, V) = K(v, v_\ell)[v_\ell - v].$$

The region $\mathbb{R}_S$ represents the synchronized state of traffic flow. The lower boundary $h_s(v)$ of this region describes the minimal distance between cars treated by drivers as safe. So if the headway distance h becomes less than $h_s(v)$ (region $\mathbb{R}_J$), the driver tries to increase it as soon as possible, which is characterized by the deceleration:

$$a_{\text{opt}}(h, v, V) = A_s^{(g)}(v_\ell)[h/T_s - v] + K_s(v_\ell)[v_\ell - v].$$

Here K , A , $A_s^{(g)}$, K_s , and T_s are some model parameters depending, in general, on v and h as well as the speed v_ℓ of the car ahead.

The ATD model also allows for the delay in driver response which is taken into account in the form of the dynamical equation governing time variations in the car acceleration:

$$\tau_{\text{del}} \frac{da}{dt} = a_{\text{opt}}(h, v, v_\ell) - a, \quad (11)$$

where τ_{del} is the delay time.

The given presentation has been confined to a simplified version of the Kerner-Klenov ATD deterministic microscopic model to focus attention on its basic features. For a detailed description of the three-phase theory of traffic flow, in particular, its microlevel representation, a reader may be referred to books by Kerner (2009a, 2017).

Summarizing the aforesaid, *first*, the car-following models outlined in the previous section are based on the common premise. Namely, the driver behavior is assumed to be governed by two stimuli which in mathematical terms are described by the relative velocity $v - v_\ell$ and the deference $V(h) - v$ representing the driver intention of keeping up a certain optimal headway distance.

Second, the fundamental difference between these models and the model based on the three-phase theory of traffic flow (the present section) is that the latter does not consider the point $v = v_\ell$, $V(h) = v$ equilibrium in synchronized flow. Instead, it turns to the concept of synchronized phase of traffic flow matching a certain 2D region, e.g., on the headway-speed plane, where each particular state can play the role of stationary point and such states make up a volumetric multitude.

Third, in this aspect, the three-phase theory of car-following and the dynamical trap model for car-following to be described below in detail are based on similar premises. Namely, the former is based on Kerner's hypotheses about the 2D region of states of synchronized flow (Kerner 1998, 1999) discussed above, whereas the latter turns to the general concept of dynamical trap (Lubashevsky 2012, 2016) also admitting the interpretation in terms of fuzzy

headway-speed diagram (see Fig. 5). The existence of a certain volumetric multitude of traffic flow long-lived states – with sharp or fuzzy boundaries – takes the central place in both the approaches.

Car-Following: Driving Simulator Experiments

The main goal of experiments on car-following reported here was:

- To demonstrate that the driver behavior has to be categorized as human intermittent control and for describing the car-following the two-component approach can be efficiently employed
- To elucidate the components of extended phase space required for describing the dynamics of car-following in a closed way

To make the following analysis of car-driving lucid, it is necessary to note one of the main results of experiments on balancing virtual pendulums with over-damped dynamics (Zgonnikov et al. 2014; Lubashevsky 2017). Due to the over-damping, the velocity v of cart moved by participants via the computer mouse for balancing the pendulum is the *main control parameter* in this case. The interpretation of the cart velocity v as the main control parameter implies that in balancing these over-damped pendulums, the cart acceleration *on its own* does not affect essentially the balancing. As a consequence, after returning the pendulum to the desired upright position, participants can just stop the cart.

In the conducted experiments on balancing virtual pendulums, their position and velocity as well as the velocity of cart were continuously recorded. Then based on the collected data, in particular, the histograms – the probability distributions – of these variables were constructed. Figure 7 depicts the found distribution $p(v)$ of the cart velocity v averaged over all the participants in the following way. First, the cart velocity distributions $\{p_i(v)\}$ were constructed individually for each participant

i using the cart velocity data normalized to the corresponding *individual* root mean square σ_{vi} of car velocity, $\sigma_{vi} = \sqrt{\langle v_i^2 \rangle_v}$. It turned out that these distributions are rather close to one another within the same setup. Then the obtained distributions were averaged over all the participants,

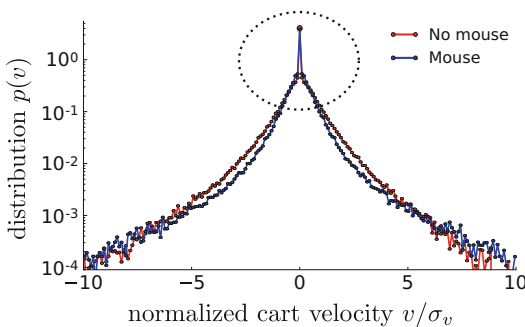
$$p(v) = \langle p_i(v/\sigma_{vi}) \rangle_i.$$

The obtained results demonstrate that the distribution function of the subjective variable controlled directly by humans in governing the system motion contains a sharp peak at the origin (Fig. 7). Exactly the intermittency in human actions is responsible for the appearance of this peak representing the relative volume of the passive phase. Its appearance:

- May be treated as the characteristic property of human intermittent behavior.
- Points out that the corresponding quantity should be treated as the main subjective parameter via which a person governs the system motion.

Experimental Setup

A car-driving simulator based on the open source engine TORCS (Access on Feb. 2018) was employed in the experiments. A standard desktop computer with a monitor located at distance about 70 cm from participant's eyes was used; the interaction between the computer and participants was



Physics of Mind and Car-Following Problem, Fig. 7 The distribution of cart velocity averaged over all the participants; their individual distributions are rather close to the shown one, based on the data collected in experiments using simulators with and without visible mouse pointer (Zgonnikov et al. 2014, 2015)

implemented via the high-precision steering wheel with pedal set (Logitech G27 Racing Wheel). The parameters of a virtual car driven by participants were set to imitate the basic dynamical characteristics typical for real vehicles of subcompact and compact car classes. A special track of rectangular form (with smoothed corners) was constructed for these experiments; its longest straight parts are of length about 70 km, and the road is of width of 15 m. The roadside of the track includes a special pattern of stripes enabling participants to perceive the car speed directly (Fig. 8).

The experiments were implemented within the car-following scenario – the participants were instructed (i) to follow the lead car without overtaking it and (ii) to keep a certain headway distance not to lose sight of the lead car. No other instructions influencing the individual preferences in car driving were given. A screenshot in Fig. 8 illustrates the car-following experiments. The set of car-following experiments consisted of trials where the lead car speed was set equal to $V = 60$ km/h, 80 km/h, 100 km/h, and 120 km/h. Each of these trials was continued for 60 min totally with possible breaks. Eight male students of age around 22–25 participated in these experiments. Other details of conducted experiments are given in Ref. (Lubashevsky and Ando 2016).

Extended Phase Space

There are at least two reasons to regard the *time derivative of car pedal position* as the desired



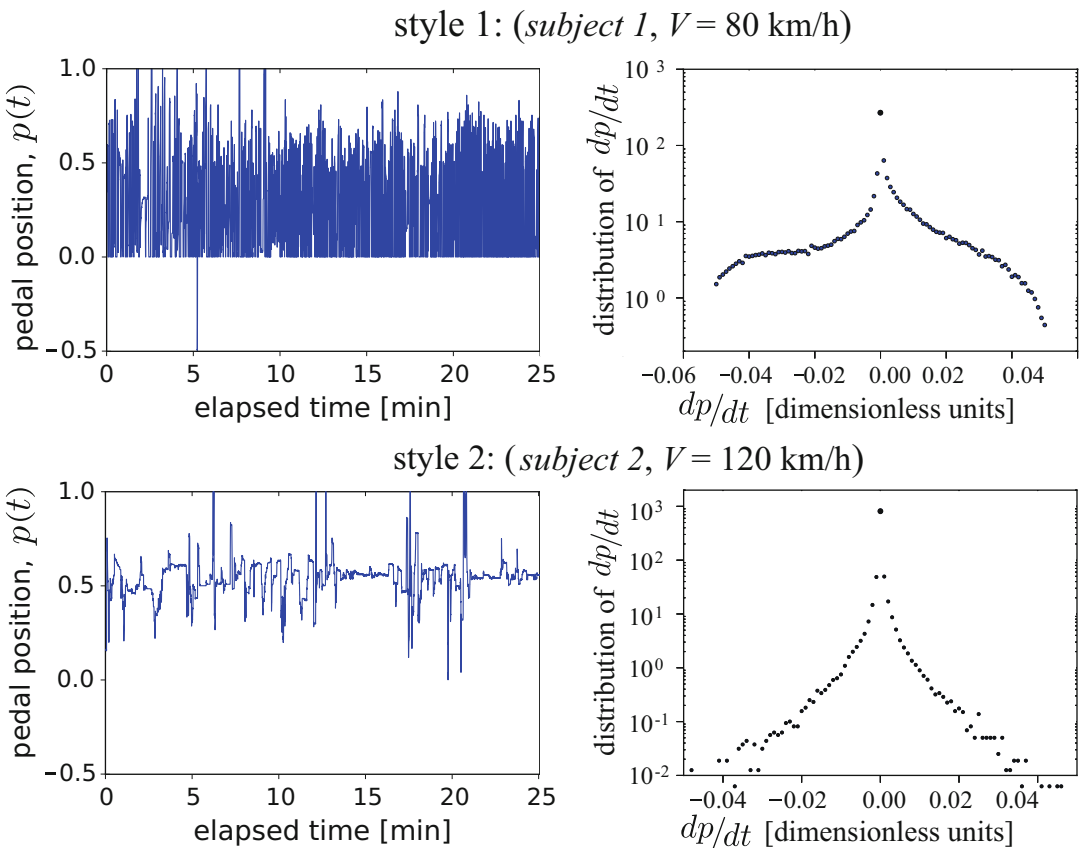
Physics of Mind and Car-Following Problem, Fig. 8 Screenshot of car-following experiments. The car ahead (in blue) is driven by computer at a fixed speed

main control parameter. The term *car pedal* denotes both the accelerator and brake pedals treated as one control unit because under the normal conditions, drivers never press both the pedals.

First, the pedal position is the quantity controlled *directly* by drivers, the time variations of other quantities – the car position, velocity, acceleration, etc. – are determined by car mechanics and the pedal position. In this sense, they are controlled by drivers indirectly. Here we ignore time scales where, e.g., the effect of human leg inertia is essential. Therefore the pedal position is a control parameter; however it cannot be regarded as the *main* control parameter. Indeed when a driver decides to keep the pedal at an arbitrary chosen position, this state of car-driving may be treated as the passive phase during which he can analyze whether the further correction of car motion is

necessary. After making decision on correcting the current state of car motion, the driver presses or releases the pedal moving it to a new position. So none of the possible values of pedal position can be attributed to the passive phase as its characteristic property. In contrast the time derivative of pedal position (*i*) characterizes the driver's actions in correcting the car motion – the phase of active behavior – and (*ii*) its zero-value matches the passive phase in driver behavior.

Second, turning to our causal experience of car-driving, we see that drivers focus their attention on the car arrangement in traffic flow, the current velocity, and acceleration rather than on the particular position of the pressed pedal. However, after making decision on correcting the current state of car motion, a driver consciously (or automatically based on the gained experience) slow or fast pushes or releases the pedal depending on the current



Physics of Mind and Car-Following Problem, Fig. 9 Characteristic forms of the time patterns of pedal position, $p(t)$ (in units of maximal shift), used in the

classification of driving styles and the corresponding distributions of the time derivative dp/dt based on Lubashevsky and Ando (2016) and Lubashevsky (2017)

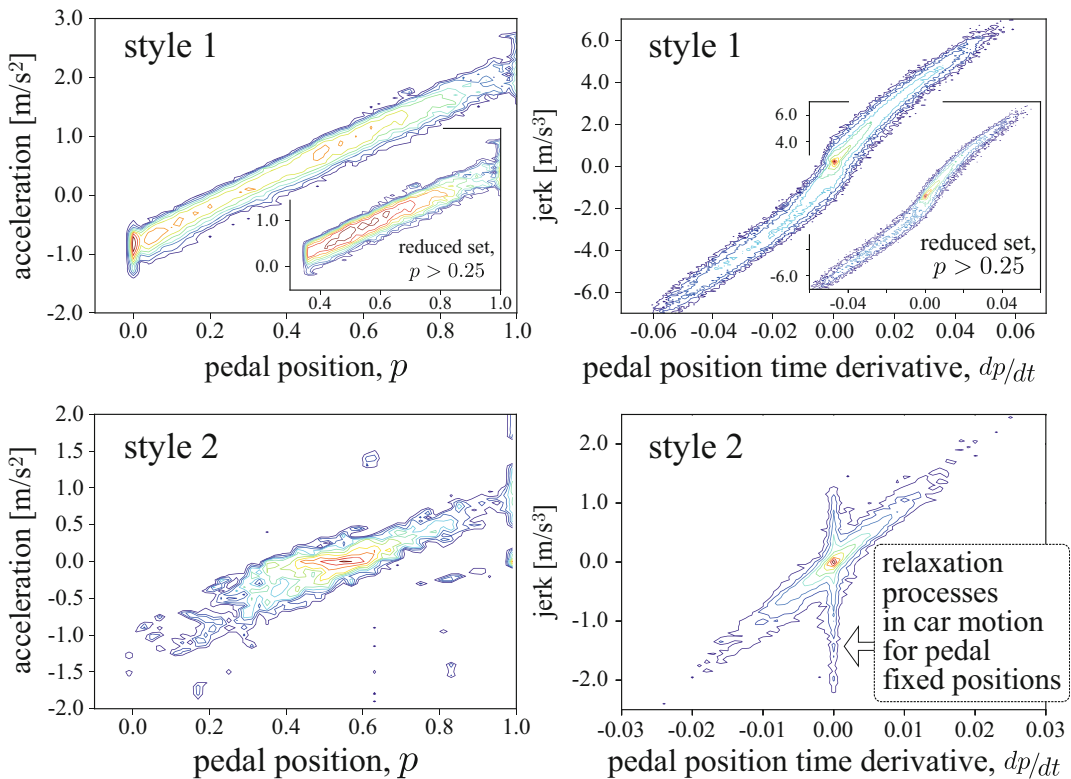
situation. In other words, the speed of pedal movement matters for car-driving and is partly determined by the safety requirements.

Figure 9 (left column) shows typical forms of the pedal position time patterns found in the conducted experiments. The time pattern classified as style 1 demonstrates the strategy of car-driving when a driver pushes the pedal for a relatively short time and then releases it also for a short time and so on. This style enables a driver to keep up the desired velocity and headway without the precise control over the pedal position just changing the duration of pressing or releasing the pedal. The time pattern classified as style 2 demonstrates the opposite strategy of car-driving, when a driver is able to keep up the required pedal position for a relatively long time interval. There is also another style of driving that may be classified as a certain mixture of the two

basic styles. The notion of *mesolevel intermittency* may be introduced to characterize this style of driving because it can be conceived of as alternate, irregular transitions between style 1 and style 2 (Lubashevsky 2017). Here our discussion is confined to the two basic styles only.

Figure 9 (right column) depicts also the corresponding distributions of the pedal position time derivative dp/dt . As shown both of them possess a common feature – a sharp peak at the origin. This result may be regarded as experimental evidence for the pedal position time derivative being the main control parameter in the car-following.

The pedal position and its time derivative characterize driver actions, whereas the explicit description of car dynamics requires another type variables such as the position of car on



Physics of Mind and Car-Following Problem, Fig. 10 The system state distribution on the phase planes {car acceleration a -pedal position p } and {car jerk j -pedal position time derivative dp/dt } for the data set shown in Fig. 9. The level lines (ten lines) are plotted in logarithmic scale. Inserts in the style1 panels depict the same

relationship when only the data-points with $p > 0.25$ are taken into account. In this way the contribution of situations when a driver releases both the accelerator and brake pedals (rather often within style 1) is eliminated (After Lubashevsky and Ando (2016))

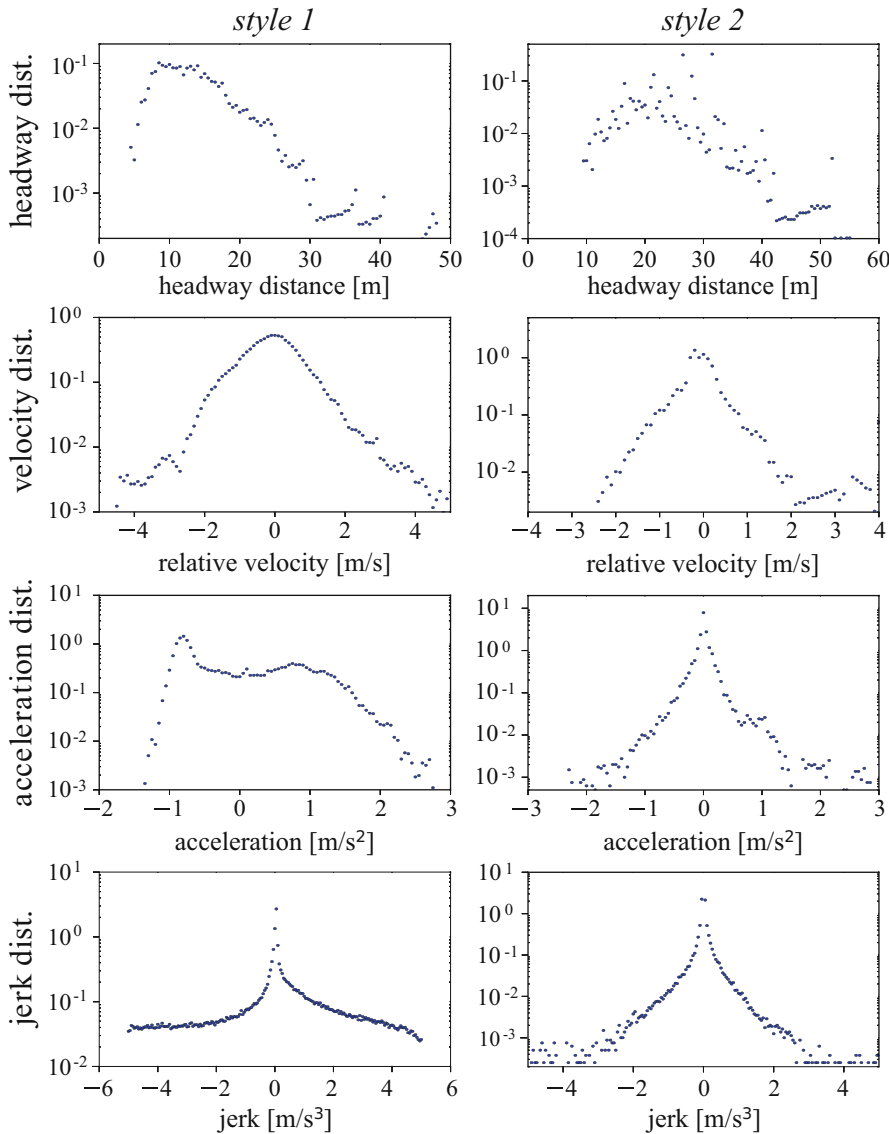
road, its velocity, acceleration, and, maybe, other quantities like the time derivative of acceleration – the jerk $j = da/dt$. Figure 10 demonstrates the relationship between the variables of car dynamics and driver actions. Based on the presented data for styles 1 and 2, the relationship between the pedal position p , its time derivative dp/dt , the car acceleration a , and the jerk j admits the linear approximation of the form (Lubashevsky 2017)

$$a = \kappa(p - pv) - \chi(v - V) \quad (12a)$$

and correspondingly when V is fixed

$$j \stackrel{\text{def}}{=} \frac{da}{dt} = \kappa \frac{dp}{dt} - \chi a. \quad (12b)$$

Here κ is some coefficient converting the pedal position into the car acceleration, the coefficient χ describes the growth in the medium resistance to



Physics of Mind and Car-Following Problem, Fig. 11 Characteristic forms of the distributions of the headway distance, the relative velocity, the car acceleration, and jerk. (After Lubashevsky and Ando (2016); Lubashevsky (2017))

car motion as the car speed increases, and p_V is the pedal position under the steady-state conditions.

Statistical properties directly characterizing the car dynamics are partly represented in Fig. 11 showing the obtained distributions of the headway distance, the car velocity, acceleration, and jerk.

First, the distribution functions of headway, car velocity, and acceleration for style 2 (Fig. 11, right column) are similar in form with respect to the corresponding distribution functions constructed based on data collected for real traffic. For example, using the GPS data collected in experiments of car-following on a test track (Gurusinghe et al. 2002) and the single car data from the German freeway A1 (Knospe et al. 2002), Wagner and Lubashevsky (2003) have found the velocity distribution to be mainly of the Laplace form $P_v \propto \exp \{-|v - v_0|/\sigma\}$, where v_0 and σ are the parameters. Using the data set of the *Intelligent Cruise Control Field Operational Test* (Fancher et al. 1998), Kendziorra et al. (2016) come to the same conclusion with respect to car acceleration. Similar results were reported by Knospe et al. (2002). The headway distribution shown in Fig. 11 is also fitted well by the popular ansätze (e.g., Luttinen 1996). These results allow us to suppose that the used experimental setup, at least, qualitatively represents human behavior in driving real cars. Naturally a more sophisticated analysis of real traffic data is necessary to elucidate this issue in detail including, in particular, the analysis of the car jerk and variations in the car pedal position.

Second, turning to Fig. 11, we see, as should be expected based on the dp/dt - j -correlation, the jerk distribution to possess a sharp peak at the origin. It enables us to regard the car jerk as an *effective* main control parameter instead of the pedal position time derivative dp/dt in describing the car-following in terms of car motion variables.

The jerk time patterns demonstrate directly typical features attributed to human intermittent control (Fig. 12). Namely, the driver actions reflected in the jerk time patterns may be conceived of as the sequence of alternate fragments matching active and passive behavior of drivers. During the passive phase, drivers do not change, at least, intentionally, the pedal position, which is clearly visible in

Fig. 12. Within the corresponding fragments of shown patterns, the acceleration changes slowly, and correspondingly, the jerk takes small values. The active phase represents driver actions on moving the pedal to a new position in correcting the car motion state; it corresponds to spike-wise variations in the variable dp/dt reflected in the shown patterns by step-wise variations in the acceleration and spike-wise variations in the jerk.

Thus, the presented results enable to state that the dynamics of car-following and the driver behavior within the analyzed conditions (Lubashevsky 2017):

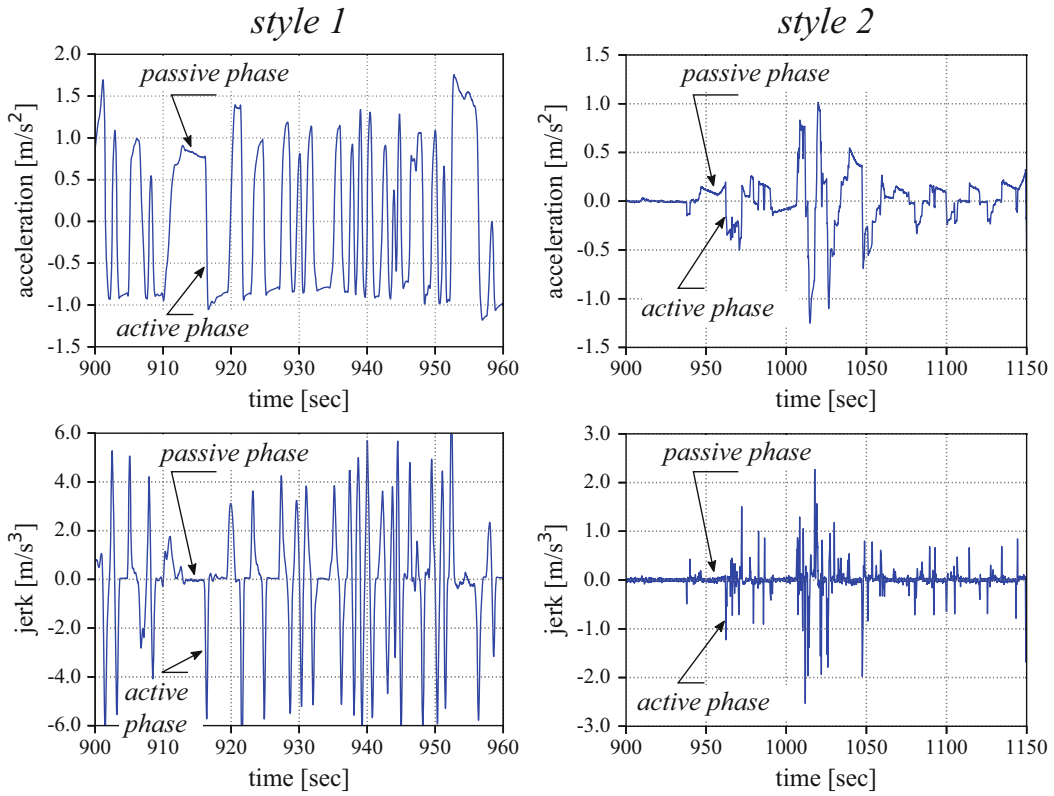
- Do exhibit the characteristic features of human intermittent behavior
- Require for their description the introduction of the extended phase space, namely, the *four-dimensional* phase space $\mathbb{P}_{cf}$ comprising the following mutually independent variables:
 - The headway distance h ,
 - The car velocity v ,
 - The car acceleration a ,
 - The pedal position time derivative dp/dt which is the main control parameter governed directly by volitional actions of drivers

Other collections of mutually independent variables, e.g., $\{h, v, a, da/dt\}$ are also possible, but the given set of phase variables has the advantage since it operates with quantities directly perceivable for drivers.

Car-Following: Dynamical Trap Model

The model for car-following presented in this section is based on (i) the results of experiments described in the previous section and (ii) the gist of dynamical trap concept outlined above. Figure 13 illustrates this model.

The dynamics of the car following the lead car which moves at a fixed speed V is described by the headway distance h between the two cars, the car velocity v , and acceleration a . These three variables make up the phase space of car motion $\mathbb{R}_c = \{h, v, a\}$. Human actions in driving the



Physics of Mind and Car-Following Problem, Fig. 12 Characteristic fragments of active phase and passive phase in the time patterns of the car acceleration and

jerk found for styles 1 and 2. (After Lubashevsky and Ando (2016); Lubashevsky (2017))

following car are described by the rate of pedal position change $\eta = dp/dt$ regarded as the main control parameter. It quantifies the driver passive and active behavior within the paradigm of human intermittent control. The phase space of car motion $\mathbb{R}_c$ and the phase space of driver behavior $\mathbb{R}_s = \{h\}$ combined together make up the four-dimensional phase space of the car-following dynamics

$$\mathbb{R}_{cf} = \mathbb{R}_c \cup \mathbb{R}_s = \left\{ h, v, a, \eta \stackrel{\text{def}}{=} \frac{dp}{dt} \right\}. \quad (13)$$

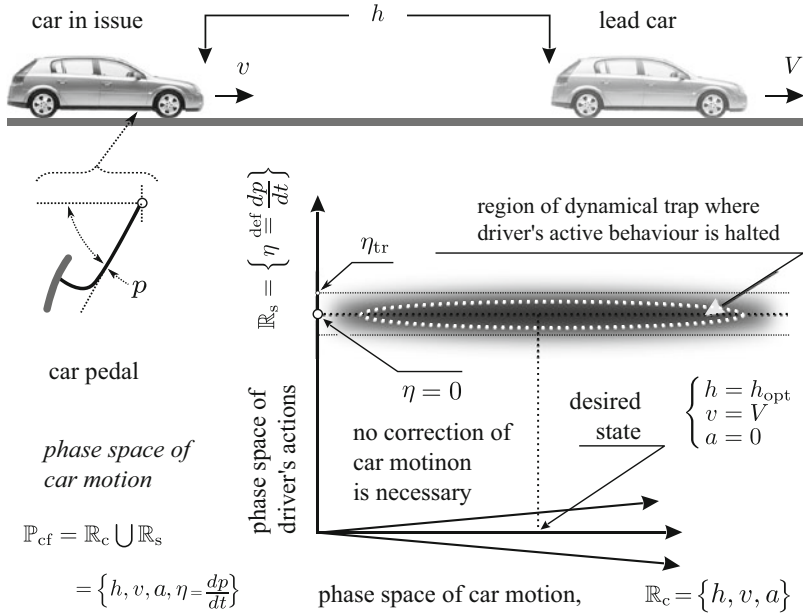
Approximation of perfectly rational drivers: if drivers were perfectly rational, then they would be able to keep up some optimal headway distance h_{opt} continuously. In order to describe the car dynamics in this case, we may

turn to the classical theory of car-following (see, e.g., car-following (see, e.g., Helbing 2001; Treiber and Kesting 2013)). In particular, the optimal headway h_{opt} can be specified by the equality $V = v_{\text{opt}}(h = h_{\text{opt}})$, where $v_{\text{opt}}(h)$ is the mean speed chosen by drivers in the real traffic flow when the mean distance between the cars is h . Let us make use of the ansatz

$$v_{\text{opt}}(h) = v_{\text{max}} \frac{h^2}{h^2 + D^2} \quad (14)$$

widely met in the traffic flow theory. Here v_{max} and D are given parameters whose values can be extracted from traffic flow data. When for some reason the car deviates from this optimal headway, the driver restores it via pressing or releasing the

Physics of Mind and Car-Following Problem, Fig. 13 The analysis of the car-following problem and the phase-space structure of the corresponding car dynamics



pedal such that the car starts to accelerate or decelerate. Turning to the well-known optimal velocity model (Bando et al. 1995), let us describe this transient process by the equation

$$\frac{dv}{dt} = a_{\text{opt}}(h, v) \stackrel{\text{def}}{=} \frac{1}{\tau_v} [v_{\text{opt}}(h) - v], \quad (15)$$

where the time scale τ_v characterizes the driver reaction delay as well as physical properties of car itself.

Dynamical trap model: In order to describe the effect of complex present on human actions in controlling the car motion by pressing or releasing the car pedal, the following dynamical trap model is used (Lubashevsky 2017):

$$\tau_s \frac{d\eta}{dt} = \Omega(\eta) \gamma [a_{\text{opt}}(h, v) - a] - \eta + \varepsilon \xi(t). \quad (16)$$

Here, according to the concept of action dynamical trap (Zgonnikov and Lubashevsky 2014), the difference $[a_{\text{opt}}(h, v) - a]$ is regarded as a stimulus to correct the current state of car dynamics described as a point of the car motion phase space $\mathbb{R}_c\{h, v, a\}$, the parameter γ is the

sensitivity of human response to the given stimulus, and the time scale τ_s evaluates the mean time of human reaction delay. The cofactor

$$\Omega(\eta) = \frac{\eta^2}{\eta^2 + \eta_{\text{tr}}^2} \quad (17)$$

represents the driver control suppression in the region of dynamical trap (Fig. 13) whose fuzzy thickness is estimated by the value η_{tr} . The last term in Eq. (16) – the Langevin force of intensity ε – allows for random factors in human actions.

Using relationship (12b) between the variables η , a , and $j = da/dt$ and converting to the dimensionless variables and parameters,

$$\begin{aligned} h &\rightarrow \frac{h}{D}, v \rightarrow \frac{v}{V}, a \rightarrow \frac{Da}{V^2}, \eta \rightarrow \frac{\kappa D^2 \eta}{V^3}, t \rightarrow \frac{Vt}{D}, \\ \varkappa &\rightarrow \frac{\varkappa D}{V}, \sigma_s = \frac{D}{V\tau_s}, \gamma \rightarrow \frac{\kappa D \gamma}{V}, \eta_{\text{tr}} \rightarrow \frac{\kappa D^2 \eta_{\text{tr}}}{V^3}, \\ \varepsilon &\rightarrow \frac{\kappa D^{5/2} \varepsilon}{V^{7/2} \tau_s}. \end{aligned}$$

the given model of car-following can be written as

$$\frac{dh}{dt} = 1 - v, \quad \frac{dv}{dt} = a, \quad \frac{da}{dt} = \eta - va, \quad (18a)$$

$$\frac{d\eta}{dt} = \frac{\eta^2}{\eta^2 + \eta_{tr}^2} r\sigma_s s(h, v, a) - \sigma_s \eta + \varepsilon \zeta(t). \quad (18b)$$

Here the function $S(h, v, a)$ quantifying the stimulus to correct the current car motion is of the form

$$s(h, v, a) = \sigma_v \left(g \frac{h^2}{h^2 + 1} - v \right) - a, \quad (19)$$

where the parameters entering formula (19) have been defined as

$$\sigma_v = \frac{D}{V\tau_v} \text{ and } g = \frac{v_{\max}}{V}. \quad (20)$$

It should be noted that the set of Eq. (18a) is related to physics of car motion as a mechanical object, whereas Eq. (18b) describes human actions in governing the car motion.

In this sense Eqs. (18a) and (18b) may be regarded as the complementary parts making up the developed theory of car-following. The former ones govern the dynamics of the system objective component, whereas the latter governs the individual dynamics of the subjective component. Both the components cannot be reduced to each other; they are governed by their own laws, and their interaction endows the system as a whole with complex properties.

Simulation and results: Before discussing the properties of system dynamics governed by the developed model (18), let us note that outside the region of dynamical trap, the system motion should be stable. It implies that the driver is able to steer the system towards the desired position $h = h_{\text{opt}}$ and $v = V$ while the system deviation from it is perceptible for him. Otherwise the system control becomes impossible in principle. In order to meet this requirement, the condition (Lubashevsky 2017)

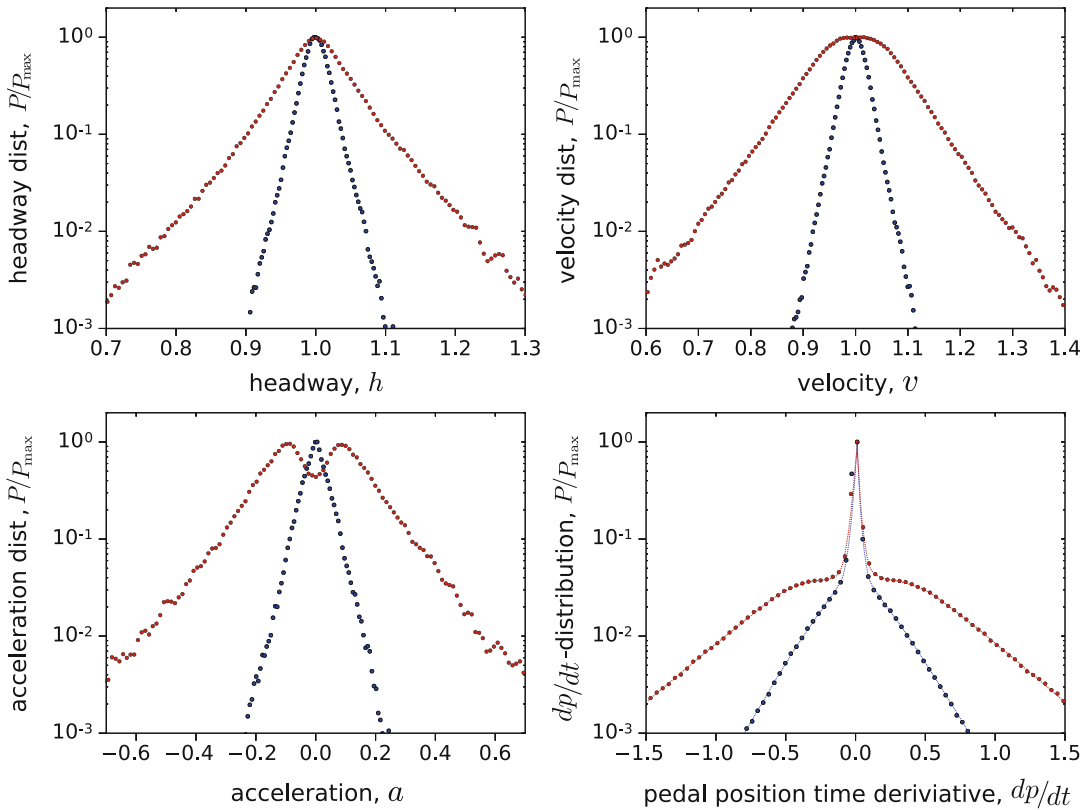
$$\gamma \left(1 - \frac{\sigma_v}{\sigma_s} \right) > \frac{2(g-1)}{h_{\text{opt}}} \quad (21)$$

should be imposed on the model parameters. Actually this condition specifies the acceptable values of the human reaction delay.

The choice of system parameter used in studying model (18) numerically is explained in Ref. (Lubashevsky 2017); here we shortly outline it. First, for typical values of $v_{\max} \sim 30$ m/s (about 120 km/h), $D \sim 15$ m, and $\tau_v \sim 1$ s used in the traffic flow theory (e.g., Treiber and Kesting 2013) and the lead car speed set equal to $V = v_{\max}/2$, we have the estimates of $\sigma_v = 1$ and $g = 2$. Second, we set $\kappa = 0.1$ which reflects weak dependence of medium resistance to car motion on the car velocity variations. Third, the estimate $\gamma \sim \sigma_s$ is justified turning to the representation of active phase of driver actions as open-loop control. Fourth, $\eta_{\text{th}} \sim \varepsilon \ll 1$ stems from the relationship between the noise intensity ε and the thickness of dynamical trap region η_{th} being actually different aspects of the human perception uncertainty. Besides, the proportionality coefficients in the $\gamma \sim \sigma_s$ and $\eta_{\text{th}} \sim \varepsilon$ as well as an estimate for ε were figured out by comparing, at least qualitatively, the results of experiments and numerical simulation. In this way, the estimates $\gamma = 2\sigma_s$, $\eta_{\text{th}}^2 = 3\varepsilon^2$, and $\varepsilon = 0.05$ have been obtained, and the system properties are compared for two particular values of σ_s , namely, $\sigma_s = 3.5$ and $\sigma_s = 6$.

Figure 14 exhibits the obtained distributions of the phase variables h , v , a , and η . As shown, the headway distributions and the velocity distributions are approximately of the same form for the given values of the parameter σ_s quantifying the driver response delay. These distributions differ from each other mainly in their width being sensitive to σ_s . Distribution functions of headway distance and car velocity obtained for styles 1 and 2 of car-driving (Fig. 11) are also of a similar form at least qualitatively.

In contrast, the form of the acceleration distribution and the distribution of the pedal position time derivative η depends essentially on the value of σ_s . In the case of $\sigma_s = 3.5$, the car acceleration is characterized by bimodal distribution, whereas the η -distribution may be conceived of as the composition of two entirely distinct components: very wide distribution with weakly pronounced maximum and a



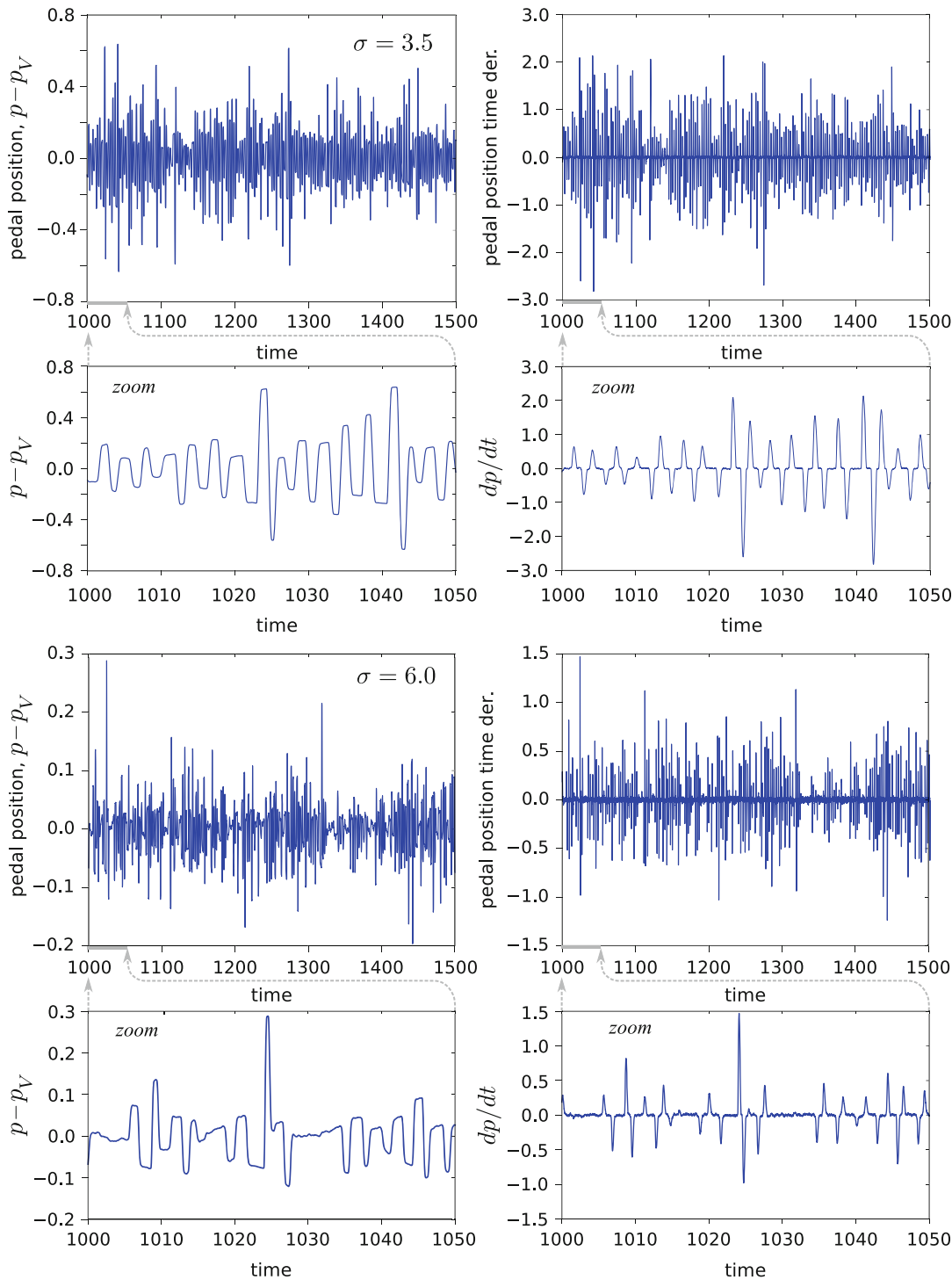
Physics of Mind and Car-Following Problem, Fig. 14 The distributions of the headway distance h , the car velocity $u = v - V_c$ and acceleration a as well as the

pedal position time derivative $\eta = dp/dt$ obtained by solving dimensionless model (18) numerically. The used parameters are noted in text. (After Lubashevsky (2017))

sharp peak at the origin. In these aspects, the given case and style 1 are similar. For $\sigma_c = 6$, the corresponding distributions obtained numerically and based on the experimental data for style 2 are also similar. The acceleration distributions are characterized by a pronounced cusp; the η -distribution contains a sharp peak at the origin, whereas its remaining part can be treated as heavy tails.

Figure 15 illustrates the time patterns obtained numerically for the pedal position ($p - p_v$) and its time derivative $\eta = dp/dt$ presented in the dimensionless units. As shown, these patterns admit the interpretation of human intermittent actions. It is worthy of noting that in

the case of $\sigma_s = 6.0$, the shown time patterns look much more irregular in comparison with ones obtained for $\sigma_s = 3.5$. Moreover, their irregularity enables us to interpret these patterns as the sequence of alternate fragments of two types. The first type fragments correspond to intensive oscillations of the system near the equilibrium position and individually comprise many such oscillations. Within the second type fragments, the system is located inside the region of dynamical trap, so time variations in the variable η are depressed on relatively long time scales. These results argue for the hypothesis about the *mesolevel* intermittency in human actions posed by Lubashevsky (2017).



Physics of Mind and Car-Following Problem, Fig. 15 The time patterns of the pedal position ($p - p_V$) and its time derivative dp/dt generated by model

(18) solved numerically. The shown variables are measured in dimensionless units, and the parameters used in simulation are noted in text. (After Lubashevsky (2017))

Conclusion

The car-following problem, on one hand, is a paradigmatic example of systems whose dynamics is governed by fundamental laws of human perception of space and motion. On the other hand, the car-following is one of the basic “bricks” making up physics of traffic flow studying various emergent phenomena arising in congested traffic on highways and road-street networks of cities. These phenomena are associated with many engineering applications important for our everyday life. Therefore trying to penetrate deeper in understanding the particular regularities of car-following, we (i) can elucidate the general mechanisms via which the human mind copes with events where spatiotemporal aspects are crucial and (ii) find an answer to challenging problems met nowadays in modeling nonlinear phenomena in traffic flow.

In *summary* to the present entry, in order to allow for the basic properties of human cognition, the description of car-following requires:

- The introduction of two complementary components – objective and subjective ones – interacting with each other. The objective component represents the mechanical aspects of car dynamics and obeys the laws of classical physics (Newtonian mechanics). The subjective component represents the driver cognition and mental evaluation of the spatial arrangement of the cars – the car he drives and the surrounding cars – as well as their current motion. The subjective component obeys its own laws irreducible to the laws of classical physics; in particular, the laws of subjective component operate with quantities that cannot be determined via referring only to the current instant of time. Even dealing with the continuous process of car-following when the long-term effects of memory and anticipation are ignorable, the notion of complex present having finite duration and some temporal structure should be used to characterize the driver response to the current stimuli.
- The introduction of the four-dimensional extended phase space comprising the objective and subjective variables. Namely, the required phase variables are the headway distance h , the velocity v , and acceleration a of the following car, as well as the rate $\eta = dp/dt$ of car pedal position change $p(t)$. The latter is the main control parameter via which the driver affects the car motion. The driver volitional actions admit the interpretation as transitions between different states of car motion and are determined by short-term anticipation within open-loop control.
- The introduction of the concept of dynamical trap – a certain fuzzy region in the extended phase space – which generalizes the notion of stationary (equilibrium) point in the car-following. The dynamical trap describes the stagnation of driver control when the car motion state is in a close proximity to the desired one. The latter can match a quasi-equilibrium configuration of car-following as well as the currently accepted strategy of correcting the car motion until the planned actions are implemented completely. The concept of dynamical trap corresponds to the special version of space-time clouds – the composition of complex present and fuzzy spatial points being the basic elements in human perception of space-time continuum – that represents the stationary point of car-following within Newtonian-type description.

Comparing (i) the presented dynamical trap model for car-following, (ii) the three-phase theory of traffic flow, and (iii) the motor behavior description based on the uncontrolled manifold, it is possible to say that:

- The three theoretical approaches accept (directly or tacitly) the concept of human fuzzy perception of spatial points and time instants. This concept endows the corresponding mathematical formalism with some regions any point of which is regarded as a desirable state in governing the system dynamics.
- The three-phase theory of traffic flow turns to the accumulated experimental and empirical data, whereas the dynamical trap model of car-following turns to the bounded capacity

of human cognition – a general concepts of human perception. As a result both of them come to the conclusion about the existence of the volumetric multitude of long-lived states, which may be regarded as an argument for their premises.

Recent investigations (Jiang et al. 2014, 2015; Tian et al. 2016) of the oscillating patterns emerging in congested traffic have revealed that the standard deviation of the vehicle velocity increases in a concave way along the platoon. This behavior of repeated deceleration-acceleration cycles in car motion is not a result of bottleneck effect or lane-changing maneuvers; also it cannot be explained as the spatial structure of *developing* instability in the homogeneous state of traffic flow. According to Jiang et al. (2014, 2015) and Tian et al. (2016), the given phenomenon can be explained turning to Kerner's hypotheses that traffic states dynamically span a 2D region on the headway-speed plane and exactly the properties of this region are responsible for the concave growth pattern of traffic oscillations. To verify whether the proposed dynamical trap model for car-following also dealing with similar 2D regions can reproduce such oscillatory patterns, it should be generalized to car ensembles. Nevertheless turning to the chain of oscillators with dynamical traps – another system with a continuum of long-lived states – we see that the dynamical trap effect can give rise to complex oscillatory patterns, dynamic phases, and order-disorder transitions (for a review, see Lubashevsky 2016, 2017). It is essential that these emergent phenomena are not a result of the classical instability of system homogeneous state but are due to the 2D structure of dynamical trap region.

Future Directions

In the presented model for car-following, the probabilistic factors in human behavior endowing the car motion with irregular dynamics are taken into account in terms of additive noise affecting the system motion in the region of dynamical trap. It

matches the novel concept of noise activation of human intermittent control (Zgonnikov et al. 2014). Nevertheless this approach only mimics the real uncertainty in human decision-making on activating or halting the control over the car motion.

The real moments of driver active behavior are events when he decides to activate or halt the control; the notion of active points describes these events. During the system motion between these actions points, the driver mainly accumulates the information about car dynamics without changing its current state.

Therefore the next step in developing the present approach to modeling car-following should be devoted to the mathematical formalism for human decision-making as action points with probabilistic properties. Its constituent elements may be outlined as follows.

- The strategies of human actions in driving a car are represented as short-term trajectories in some extended phase space allowing for human anticipation. Broadly speaking, these trajectories are the basic objects the complex present deals with.
- Human uncertainty in perceiving and evaluating such trajectories can be quantified within the Lagrangian formalism with higher-order time derivatives. These trajectories have to be treated as some points of the space of possible strategies in car-driving, and the notion of dynamical trap should be introduced for the given space of possible strategies.
- In car-driving the decision-making may be conceived of as the accumulation of evidence for selecting one or another strategy of behavior. In particular, it requires endowing the subjective component with an additional dimension allowing for the evidence accumulation as an internal (mental) process. In this case, the action points are represented as the moments of decision-making based on the accumulated information.

As far as the relation between the three-phase theory of traffic flow and the dynamical trap model for car-following is concerned:

- Relatively fast variations in the car arrangement during the active correction of car position in traffic flow may enable the introduction of effective sharp boundaries of the multitude of long-lived states. In this case, the dynamical trap region admits interpretation in terms of the 2D region of possible metastable state introduced within the three-phase theory.
- It opens a gate toward developing more general models for traffic flow turning to the available experimental data and the basic concepts of physics of the human mind.

Summarizing the aforesaid, we can describe the further steps in developing the dynamical trap model for car-following as constructing the probabilistic theory of action points treated as random transitions in the space of possible strategies of car-driving. It should be emphasized that the development of such a theory can open a gate to describing general phenomena governed by human intermittent behavior.

Bibliography

- Andersen HK, Grush R (2009) A brief history of time-consciousness: historical precursors to James and Husserl. *J Hist Philos* 47(2):277–307
- Andrews K (2015) *The animal mind: an introduction to the philosophy of animal cognition*. Routledge, Taylor & Francis Group, London
- Anonymous (E Robert Kelly) (1882) *The alternative: a study in psychology*. Macmillan and Co., London
- Asai Y, Tasaka Y, Nomura K, Nomura T, Casadio M, Morasso P (2009) A model of postural control in quiet standing: robust compensation of delay-induced instability using intermittent activation of feedback control. *PLoS One* 4(7):e6169 (1–14)
- Asai Y, Tateyama S, Nomura T (2013) Learning an intermittent control strategy for postural balancing using an EMG-based human-computer interface. *PLoS One* 8(5):e62, 956 (19 pages)
- Balasubramaniam R (2013) On the control of unstable objects: the dynamics of human stick balancing. In: Richardson M, Riley M, Shockley K (eds) *Progress in motor control: neural, computational and dynamic approaches*. Springer Science + Business Media, New York, pp 149–168
- Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51:10351042
- Bando M, Hasebe K, Nakanishi K, Nakayama A (1998) Analysis of optimal velocity model with explicit delay. *Phys Rev E* 58:5429–5435
- Bernstein NA (1935) The problem of interrelation between coordination and localization. *Arch Biol Sci* 38:1–35, (in Russian)
- Bifulco GN, Pariota L, Brackstone M, McDonald M (2013) Driving behaviour models enabling the simulation of advanced driving assistance systems: revisiting the action point paradigm. *Transp Res C Emerg Technol* 36:352–366. <https://doi.org/10.1016/j.trc.2013.09.009>
- Bottaro A, Yasutake Y, Nomura T, Casadio M, Morasso P (2008) Bounded stability of the quiet standing posture: an intermittent control model. *Hum Mov Sci* 27(3):473–495
- Bourne C (2006) *A future for presentism*. Oxford University Press, Oxford, UK
- Brackstone M, McDonald M (1999) Car-following: a historical review. *Transport Res F Traffic Psychol Behav* 2(4):181–196
- Carter P, Christiansen PL, Gaididei YB, Gorria C, Sandstede B, Sørensen MP, Starke J (2014) Multijam solutions in traffic models with velocity-dependent driver strategies. *SIAM J Appl Math* 74(6):1895–1918
- Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. *Rev Mod Phys* 81(2):591–646
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329(4–6):199–329
- Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci* 36(03):181–204. <https://doi.org/10.1017/S0140525X12000477>
- Edie LC (1961) Car-following and steady-state theory for noncongested traffic. *Oper Res* 9(1):66–76. <https://doi.org/10.1287/opre.9.1.66>
- Elliott MA, Giersch A (2016) What happens in a moment. *Front Psychol* 6:Article 1905 (7 pages). <https://doi.org/10.3389/fpsyg.2015.01905>
- Fancher P, Haugen J, Buonarosa ML, Bareket Z, Bogard SE, Hagan MR, Sayer JR, Ervin RD (1998) *Intelligent cruise control field operational test. Final report. Volume I: Technical report. Report number: UMTRI-98-17, Report number: DOT/HS 808 849*. University of Michigan, Transportation Research Institute, Ann Arbor
- Friston K, Rigoli F, Ognibene D, Mathys C, Fitzgerald T, Pezzulo G (2015) Active inference and epistemic value. *Cogn Neurosci* 6(4):187–214. <https://doi.org/10.1080/17588928.2015.1020053>
- Gaididei YB, Gorria C, Berkemer R, Kawamoto A, Shiga T, Christiansen PL, Sørensen MP, Starke J (2013) Controlling traffic jams by time modulating the safety distance. *Phys Rev E* 88(4):042,803
- Gallagher S, Zahavi D (2012) *The phenomenological mind: an introduction to philosophy of mind and cognitive science*, 2nd edn. Routledge, Taylor & Francis Group, London

- Gawthrop P, Loram I, Lakie M, Gollee H (2011) Intermittent control: a computational theory of human control. *Biol Cybern* 104(1–2):31–51
- Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady-state traffic flow. *Oper Res* 7(4):499–505. <https://doi.org/10.1287/opre.7.4.499>
- Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9(4):545–567
- Gurusinghe GS, Nakatsuji T, Azuta Y, Ranjitkar P, Tanaboriboon Y (2002) Multiple car-following data with real-time kinematic global positioning system. *Transp Res Rec J Transp Res Board* 1802(1):166–180
- Haken H (2009a) Synergetics: basic concepts. In: Meyers RA (ed) *Encyclopedia of complexity and systems science*. Springer Science+Business Media, New York, pp 8926–8946
- Haken H (2009b) Synergetics, introduction to. In: Meyers RA (ed) *Encyclopedia of complexity and systems science*. Springer Science+Business Media, New York, pp 8946–8948
- Hawley K (2015) Temporal parts In: Zalta EN (ed) *The Stanford encyclopedia of philosophy*, winter 2015 edn. Metaphysics Research Lab, Stanford University, Stanford, CA
- Helbing D (1991) A mathematical model for the behavior of pedestrians. *Behav Sci* 36(4):298–310
- Helbing D (1994) A mathematical model for the behavior of individuals in a social field. *J Math Sociol* 19(3):189–219
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:10671141
- Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51(5):4282
- Helbing D, Tilch B (1998) Generalized force model of traffic dynamics. *Phys Rev E* 58(1):133–138. <https://doi.org/10.1103/PhysRevE.58.133>
- Hoogendoorn S, Hoogendoorn R, Daamen W (2011) Wiedemann revisited – new trajectory filtering technique and its implications for car-following modeling. *Transp Res Rec J Transp Res Board* 2260:152–162. <https://doi.org/10.3141/2260-17>
- Huys R, Perdikis D, Jirsa VK (2014) Functional architectures and structured flows on manifolds: a dynamical framework for motor behavior. *Psychol Rev* 121(3):302–336
- James W (1890) *The principles of psychology*, vol 1. Henry Holt and Company, New York
- Jiang R, Wu Q, Zhu Z (2001) Full velocity difference model for a car-following theory. *Phys Rev E* 64(1):017,101 (4 pages). <https://doi.org/10.1103/PhysRevE.64.017101>
- Jiang R, Hu MB, Zhang HM, Gao ZY, Jia B, Wu QS, Wang B, Yang M (2014) Traffic experiment reveals the nature of car-following. *PLoS One* 9(4):e94, 351 (9 pages). <https://doi.org/10.1371/journal.pone.0094351>
- Jiang R, Hu MB, Zhang H, Gao ZY, Jia B, Wu QS (2015) On some experimental features of car-following behavior and how to model them. *Transp Res B Methodol* 80:338–354. <https://doi.org/10.1016/j.trb.2015.08.003>
- Kendziorra A, Wagner P, Toledo T (2016) A stochastic car following model. *Transp Res Procedia* 15:198–207. <https://doi.org/10.1016/j.trpro.2016.06.017>, International Symposium on Enhancing Highway Performance (ISEHP), June 14–16, 2016, Berlin
- Kerner BS (1998) Experimental features of self-organization in traffic flow. *Phys Rev Lett* 81(17):3797–3800. <https://doi.org/10.1103/PhysRevLett.81.3797>
- Kerner BS (1999) Theory of congested traffic flow: self-organization without bottlenecks. In: Ceder A (ed) *Transportation and traffic theory: proceedings of the 14th international symposium on transportation and traffic theory Jerusalem, Israel, 20–23 July, 1999*. Elsevier Science, Oxford, UK, pp 147–171
- Kerner BS (2004) *The physics of traffic: empirical freeway pattern features, engineering applications, and theory*. Springer, Berlin
- Kerner BS (2009a) *Introduction to modern traffic flow theory and control: the long road to three-phase traffic theory*. Springer, Berlin
- Kerner BS (2009b) Traffic congestion, modeling approaches to. In: Meyers RA (ed) *Encyclopedia of complexity and systems science*. Springer- Science+- Business Media, New York, pp 9302–9355
- Kerner BS (2013) Criticism of generally accepted fundamentals and methodologies of traffic and transportation theory: a brief review. *Phys A Stat Mech Appl* 392(21):5261–5282. <https://doi.org/10.1016/j.physa.2013.06.004>
- Kerner BS (2016) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Phys A Stat Mech Appl* 450:700–747. <https://doi.org/10.1016/j.physa.2016.01.034>
- Kerner BS (2017) *Breakdown in traffic networks: fundamentals of transportation science*. Springer, Berlin. <https://doi.org/10.1007/978-3-662-54473-0>
- Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35(3):L31–L43. <https://doi.org/10.1088/0305-4470/35/3/102>
- Kerner BS, Klenov SL (2003) Microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68(3):036,130 (20 pages). <https://doi.org/10.1103/PhysRevE.68.036130>
- Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A Math Gen* 39(8):1775–1809. <https://doi.org/10.1088/0305-4470/39/23/C01>
- Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. *Phys Rev Lett* 79(202A-3):4030–4033. <https://doi.org/10.1103/PhysRevLett.79.4030>
- Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35(47):9971–10,013. <https://doi.org/10.1088/0305-4470/35/47/303>
- Kerner BS, Klenov SL, Hiller A (2006) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. *J Phys A Math Gen* 39(9):2001–2020. <https://doi.org/10.1088/0305-4470/39/9/002>

- Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: microscopic description of traffic phases. *Phys Rev E* 65(5):056–133 (16 pages)
- Krauss S, Wagner P, Gawron C (1996) Continuous limit of the Nagel-Schreckenberg model. *Phys Rev E* 54(4):3707–3712. <https://doi.org/10.1103/PhysRevE.54.3707>
- Latash ML (2008) Synergy. Oxford University Press, Oxford
- Latash ML (2012a) Fundamentals of motor control. Elsevier, London
- Latash ML (2012b) The bliss (not the problem) of motor abundance (not redundancy). *Exp Brain Res* 217(1):1–5. <https://doi.org/10.1007/s00221-012-3000-4>
- Latash ML (2015) Bernstein's "desired future" and physics of human movement. In: Nadin M (ed) Anticipation: learning from the past the Russian/soviet contributions to the science of anticipation. Springer, Cham, pp 287–300
- Latash ML, Scholz JP, Schöner G (2007) Toward a new theory of motor synergies. *Mot Control* 11(3):276–308
- Lewin K (1951) In: Cartwright D-w (ed) Field theory in social science: selected theoretical papers. Harpers, Oxford, UK
- Li QL, Wong S, Min J, Tian S, Wang BH (2016) A cellular automata traffic flow model considering the heterogeneity of acceleration and delay probability. *Phys A Stat Mech Appl* 456:128134. <https://doi.org/10.1016/j.physa.2016.03.026>
- Loram I, Gollee H, Lakie M, Gawthrop P (2011) Human control of an inverted pendulum: is continuous control necessary? Is intermittent control effective? Is intermittent control physiological? *J Physiol* 589(2):307–324
- Lubashevsky I (2012) Dynamical traps caused by fuzzy rationality as a new emergence mechanism. *Adv Complex Syst* 15(8):1250,045 (25 pages). <https://doi.org/10.1142/S0219525912500452>
- Lubashevsky I (2016) Human fuzzy rationality as a novel mechanism of emergent phenomena. In: Skiadas CH, Skiadas C (eds) Handbook of applications of Chaos theory. CRC Press, Taylor & Francis Group, London, pp 827–878
- Lubashevsky I (2017) Physics of the human mind. Springer International Publishing AG, Cham. <https://doi.org/10.1007/978-3-319-51706-3>
- Lubashevsky I, Ando H (2016) Intermittent control properties of car following: theory and driving simulator experiments. arXiv preprint physics 160901812, pp 125
- Lubashevsky IA, Gafiychuk VV, Demchuk AV (1998) Anomalous relaxation oscillations due to dynamical traps. *Phys A Stat Mech Appl* 255(3–4):406–414. [https://doi.org/10.1016/S0378-4371\(98\)00094-6](https://doi.org/10.1016/S0378-4371(98)00094-6)
- Lubashevsky I, Kalenkov S, Mahnke R (2002a) Towards a variational principle for motivated vehicle motion. *Phys Rev E* 65:036,140, (5 pages). <https://doi.org/10.1103/PhysRevE.65.036140>
- Lubashevsky I, Mahnke R, Wagner P, Kalenkov S (2002b) Long-lived states in synchronized traffic flow: empirical prompt and dynamical trap model. *Phys Rev E* 66:016,117, 13 pages
- Lubashevsky I, Hajimahmoodzadeh M, Katsnelson A, Wagner P (2003a) Noise-induced phase transition in an oscillatory system with dynamical traps. *Eur Phys J B Condens Matter Complex Syst* 36(1):115–118. <https://doi.org/10.1140/epjb/e2003-00323-0>
- Lubashevsky I, Wagner P, Mahnke R (2003b) Bounded rational driver models. *Eur Phys J B Condens Matter Complex Syst* 32(2):243–247. <https://doi.org/10.1140/epjb/e2003-00094-6>
- Lubashevsky I, Wagner P, Mahnke R (2003c) Rational-driver approximation in car-following theory. *Phys Rev E* 68(5):056,109, (15 pages). <https://doi.org/10.1103/PhysRevE.68.056109>
- Luttinen RT (1996) Statistical analysis of vehicle time headways. PhD thesis, Helsinki University of Technology, Lahti Cente, Neopoli, Lahti
- Maerivoet S, Moor BD (2005) Cellular automata models of road traffic. *Phys Rep* 419(1):1–4. <https://doi.org/10.1016/j.physrep.2005.08.005>
- Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408(1):1–130. <https://doi.org/10.1016/j.physrep.2004.12.001>
- Marschler C, Sieber J, Berkemer R, Kawamoto A, Starke J (2014) Implicit methods for equation-free analysis: convergence results and analysis of emergent waves in microscopic traffic models. *SIAM J Appl Dyn Syst* 13(3):1202–1238
- Marschler C, Sieber J, Hjorth PG, Starke J (2015) Equation-free analysis of macroscopic behavior in traffic and pedestrian flow. In: Chraïbi M, Boltes M, Schadschneider A, Seyfried A (eds) Traffic and granular Flow'13. Springer International Publishing, Cham, pp 423–439
- Martin V, Scholz JP, Schöner G (2009) Redundancy, self-motion, and motor control. *Neural Comput* 21(5):1371–1414
- Milton JG (2013) Intermittent motor control: the "drift-and-act" hypothesis. In: Richardson MJ, Riley MA, Shockley K (eds) Progress in motor: control neural, computational and dynamic approaches. Springer Science+Business Media, New York, pp 169–193
- Nagatani T, Nakanishi K (1998) Delay effect on phase transitions in traffic dynamics. *Phys Rev E* 57(6):6415–6421. <https://doi.org/10.1103/PhysRevE.57.6415>
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I* 2(12):2221–2229. <https://doi.org/10.1051/jp1:1992277>
- Newell GF (1961) Nonlinear effects in the dynamics of car following. *Oper Res* 9(2):209–229
- Oullier O, Kelso J (2009) Social coordination, from the perspective of coordination dynamics. In: Meyers R (ed) Encyclopedia of complexity and systems science. Springer Science +Business Media, LLC, New York, pp 8198–8213
- Pariota L, Bifulco GN (2015) Experimental evidence supporting simpler action point paradigms for car-following. *Transport Res F Traffic Psychol Behav* 35:1–15. <https://doi.org/10.1016/j.trf.2015.08.002>
- Perlovsky LI (2006) Toward physics of the mind: concepts, emotions, consciousness, and symbols. *Phys Life Rev* 3(1):23–55. <https://doi.org/10.1016/j.plrev.2005.11.003>
- Perlovsky LI (2016) Physics of the mind. *Front Syst Neurosci* 10:Article 84 (12 pages). <https://doi.org/10.3389/fnsys.2016.00084>

- Pezzulo G, Rigoli F, Friston K (2015) Active inference, homeostatic regulation and adaptive behavioural control. *Prog Neurobiol* 134:17–35. <https://doi.org/10.1016/j.pneurobio.2015.09.001>
- Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24(3):274–281
- Proctor RW, Vu KPL (2003) Action selection. In: Weiner IB, Healy AF, Proctor RW (eds) *Handbook of psychology. Experimental psychology*, vol 4. Wiley, Hoboken, pp 293–316
- Reiter U (1994) Empirical studies as basis for traffic flow models. In: Alçelik R and Reilly W (eds) *Proceedings of the second international symposium on highway capacity (Sydney, N.S.W., 1994)*, vol 2. : Australian Road Research Board, Vermont South, Victoria, Australia pp 493–502
- Reuschel A (1950a) Vehicle movements in a platoon. *Österr Ing Arch* 4:193–215
- Reuschel A (1950b) Vehicle movements in a platoon with uniform acceleration or deceleration of the lead vehicle. *Z Österr Ing Arch Ver* 95:50–62; 73–77
- Schmidt RA (1975) A schema theory of discrete motor skill learning. *Psychol Rev* 82(4):225–260
- Schmidt RA (2003) Motor schema theory after 27 years: reflections and implications for a new theory. *Res Q Exerc Sport* 74(4):366–375
- Schmidt RA, Lee TD (2011) *Motor control and learning: a behavioral emphasis*, 5th edn. Human Kinetics, Champaign
- Schoeller F, Perlovsky L, Arseniev D (2018) Physics of mind: experimental confirmations of theoretical predictions. *Phys Life Rev.* <https://doi.org/10.1016/j.plrev.2017.11.021>
- Scholz PJ, Schöner G (1999) The uncontrolled manifold concept: identifying control variables for a functional task. *Exp Brain Res* 126(3):289–306. <https://doi.org/10.1007/s002210050738>
- Smith DW (2013) Phenomenology In: Zalta EN (ed) *The Stanford encyclopedia of philosophy*, winter 2016 edn. Metaphysics Research Lab, Stanford University, Stanford, CA
- Suzuki Y, Nomura T, Casadio M, Morasso P (2012) Intermittent control with ankle, hip, and mixed strategies during quiet standing: a theoretical proposal based on a double inverted pendulum model. *J Theor Biol* 310:55–79
- Tian J, Jiang R, Jia B, Gao Z, Ma S (2016) Empirical analysis and simulation of the concave growth pattern of traffic oscillations. *Transp Res B Methodol* 93:338–354. <https://doi.org/10.1016/j.trb.2016.08.001>
- Todosiev EP (1963) The action point model of the driver-vehicle system. Technical report 202A–3, PhD dissertation, Ohio State University
- Todosiev EP, Barbosa LC (1963/64) A proposed model for the driver-vehicle system. *Traffic Eng* 34:17–20
- TORCS (2018). The official site of TORCS. <http://torcs.sourceforge.net/index.php>. Accessed on Feb 2018
- Treiber M, Helbing D (1999) Macroscopic simulation of widely scattered synchronized traffic states. *J Phys A Math Gen* 32(1):L17–L23. <https://doi.org/10.1088/0305-4470/32/1/003>
- Treiber M, Kesting A (2013) *Traffic flow dynamics: data, models and simulation*. Springer, Berlin
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62:1805–1824
- Wagner P (2006) How human drivers control their vehicle. *Eur Phys J B Condens Matter Complex Syst* 52(3):427–431. <https://doi.org/10.1140/epjb/e2006-00300-1>
- Wagner P (2012) Analyzing fluctuations in car-following. *Transp Res B Methodol* 46(10):1384–1392
- Wagner P, Lubashevsky I (2003) Empirical basis for car-following theory development. arXiv preprint condmat/0311192. <http://adsabs.harvard.edu/abs/2003cond.mat.11192W>
- Yoshikawa N, Suzuki Y, Kiyono K, Nomura T (2016) Intermittent feedback-control strategy for stabilizing inverted pendulum on manually controlled cart as analogy to human stick balancing. *Front Comput Neurosci* 10:Article 34, 19 pages. <https://doi.org/10.3389/fncom.2016.00034>
- Yu S, Liu Q, Li X (2013) Full velocity difference and acceleration model for a car-following theory. *Commun Nonlinear Sci Numer Simul* 18(5):1229–1234. <https://doi.org/10.1016/j.cnsns.2012.09.014>
- Zaslavsky GM (1995) From Hamiltonian chaos to Maxwell's demon. *Chaos* 5(4):653–661
- Zaslavsky GM (2002) Dynamical traps. *Phys D Nonlinear Phenom* 168–169:292–304. [https://doi.org/10.1016/S0167-2789\(02\)00516-X](https://doi.org/10.1016/S0167-2789(02)00516-X) {VII} Latin American Workshop on Nonlinear Phenomena
- Zaslavsky GM (2005) *Hamiltonian chaos and fractional dynamics*. Oxford University Press, New York
- Zgonnikov A, Lubashevsky I (2014) Extended phase space description of human-controlled systems dynamics. *Prog Theor Exp Phys* 2014(3):033J02
- Zgonnikov A, Lubashevsky I (2015) Double-well dynamics of noise-driven control activation in human intermittent control: the case of stick balancing. *Cogn Process* 16(4):351–358. <https://doi.org/10.1007/s10339-015-0653-5>
- Zgonnikov A, Lubashevsky I, Kanemoto S, Miyazawa T, Suzuki T (2014) To react or not to react? Intrinsic stochasticity of human control in virtual stick balancing. *J R Soc Interface* 11:20140,636
- Zgonnikov A, Kanemoto S, Lubashevsky I, Suzuki T (2015) How the type of visual feedback affects actions of human operators: the case of virtual stick balancing. In: *Systems, man, and cybernetics (SMC)*, 2015 I.-E. international conference on. IEEE, Piscataway, NJ pp 1100–1103. <https://doi.org/10.1109/SMC.2015.197>
- Zhao X, Gao Z (2005) A new car-following model: full velocity and acceleration difference model. *Eur Phys J B Condens Matter Complex Syst* 47(1):145–150. <https://doi.org/10.1140/epjb/e2005-00304-3>

Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems

Takashi Nagatani
Department of Mechanical Engineering,
Shizuoka University, Hamamatsu, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Model and Similarity
Nonlinear Map Model
Freely Passing Buses
No Passing Buses
Increase or Decrease of Buses
Future Directions
Bibliography

Glossary

Combined map model The combined map is a nonlinear map combined with two nonlinear maps. The combined map model is a dynamic model of a single bus in which both effects of speed control and periodic inflow is taken into account.

Delayed map model Nonlinear map with a time lag. The delayed map model is a dynamic model of bus motion described by the delayed map where the delay of acceleration or deceleration is taken into account.

Deterministic chaos The future behavior of dynamics involved no random elements does not make predictable. The behavior is known as deterministic chaos. The chaos occurs in the nonlinear map models of buses, trams, and elevators.

Extended circle map model The circle map is a one-dimensional map which maps a circle onto itself. A nonlinear map including sin function.

The extended circle map model is a dynamic model of a bus with the periodic inflow at a transfer point.

Nonlinear map model A nonlinear map is a recurrence relation presented by a nonlinear function. The nonlinear map model is a dynamic model of bus motion described by the nonlinear map.

Piecewise map model A recurrence relation presented by a continuous function comprised of segments. The piecewise map model is a dynamic model of bus motion described by the piecewise map.

Definition of the Subject

It is necessary and important to operate buses and trams on time. The bus schedule is closely related to the dynamic motion of buses. In this part, we introduce the nonlinear maps for describing the dynamics of shuttle buses in the transportation system. The complex motion of the buses is explained by the nonlinear map models. The transportation system of shuttle buses without passing is similar to that of the trams. The transport of elevators is also similar to that of shuttle buses which are freely passing. The complex dynamics of a single bus is described in terms of the piecewise map, the delayed map, the extended circle map, and the combined map. The dynamics of a few buses is described by the model of freely passing buses, the model of no passing buses, and the model of increase or decrease of buses. The nonlinear map models are useful to make an accurate estimate of the arrival time of the bus in the terminal.

Introduction

The traffic science aims to discover the fundamental properties and laws and the traffic engineering aims at making the planning and implementation of road networks and control systems. Physics, other sciences, and technologies meet at the

frontier area of interdisciplinary research. The traffic flow has been studied from a point of view of statistical physics and nonlinear dynamics (Kerner 2004; Treiber et al. 2000, 2006; Lämmer et al. 2006; Kerner 2016a, b, 2017; Li et al. 2017; Gupta et al. 2015; Redhu and Gupta 2016). Various models have been presented to understand the rich variety of physical phenomena exhibited by traffic. Analytical and numerical techniques have been applied to study these models. The concepts and techniques of physics can be applied to transportation systems. Specially, the traffic jams and congestion have been investigated. The vehicular traffic controlled by signals has been studied by some physical models (Toledo et al. 2004, 2007; Wastavino et al. 2008; Villalobos et al. 2010; Nagatani 2005a, 2006a, 2007a; Komada et al. 2011). Also, the dynamic models for multi-vehicle collisions have been proposed (Naito and Nagatani 2012; Sugiyama and Nagatani 2012, 2013).

Some information of traffic has an important effect on the underlying dynamics. In the real traffic, advanced traveler information systems provide real-time information about the traffic conditions to road users by means of communication. Real-time information helps the individual road users to minimize their personal travel time. The dynamic models have been proposed for the route choice of traffic system with real-time information (Wahle et al. 2000; Dong et al. 2010; Tobita and Nagatani 2012, 2013; Hino and Nagatani 2014, 2015; Nagatani 2013a, 2014). Despite the complexity of traffic, physical traffic theory is an example of a highly quantitative description for a living system.

In the public transportation system, it is important and necessary to make a schedule of public transport, for example, bus schedule. It is well known that passengers using buses are served best when buses arrive at stations on time, and there is no congestion. Frequently, buses are delayed or go faster in the bus transport system. Thus, it is hard to operate buses on time. Bus schedules are closely related to the dynamic motion of buses. The bus dynamics depends highly on the control method and the system. The arrival time of buses is not determined only by serving passengers but also by the speed control method of buses. The dynamics of

trams and elevators is also similar to that of the bus transport system.

In the morning and evening peaks, the bus transport system exhibits severe congestion problems. The maximum rate of serving passengers increases with the number of buses. In managing the bus operation, the usual criterion for deciding the number of buses is that one should be able to transport everyone from the starting point to his destination within some period of time for the rush hour trips. Another criterion used in shuttle bus operation is that a passenger's waiting time should not exceed some specified value. It is very important to make an accurate estimate of the arrival time of the bus in the terminal. There also can be the same problems in the tram and elevator systems.

Until now, some models of the bus transport system have been studied. In the bus route model with many buses, it has been found that the bunching transition between a heterogeneously jammed phase and a homogeneous phase occurs with increasing density. Such dynamic models as the car-following model have been proposed to study the bunching transition (O'loan et al. 1998; Nagatani 2000, 2001a; Chowdhury and Desai 2000; Huijberts 2002). Little is known about the travel time and delay of buses.

In shuttle bus systems, it is important to make correct bus schedules. It is necessary to calculate and derive the travel time and delay. The bus schedule must be determined by the dynamic motion of buses considering the inflow rate of passengers and the speed control method. By taking into account the speed control method and bus transport system, one should estimate the bus schedule. The dynamic model of the bus system is necessary to estimate the arrival time (Nagatani 2001b, c, 2002a, b, c, 2003a, 2008; Nagatani and Yoshimura 2002).

It is well known that the delay of acceleration and deceleration has the important effect on the jam formation in traffic flow. Many researchers have studied the delay effect for the vehicular traffic flow. However, the delay effect on bus schedule is unknown about the bus transportation system. It is important to study how the delay of speedup affects the bus dynamics and its schedule. There are some cases that the motion of buses is unstable if one retrieves the delay of buses.

Buses show the periodic, quasiperiodic, and chaotic motions. It is difficult to make bus schedule due to the complex behavior.

In this part, we present the nonlinear map models for the complex dynamics of buses, trams, and elevators. The bus schedule is derived from the nonlinear map model for the dynamic motion of a shuttle bus. We make an accurate estimate of the arrival time using the nonlinear map model. We show that both inflow rate and speedup control have an important effect on the dynamic transitions to complex motion occurring in the bus traffic. Also, we present the nonlinear map models for the transport system of a few buses. We discuss the relationship between the bus schedule and the dynamic motion.

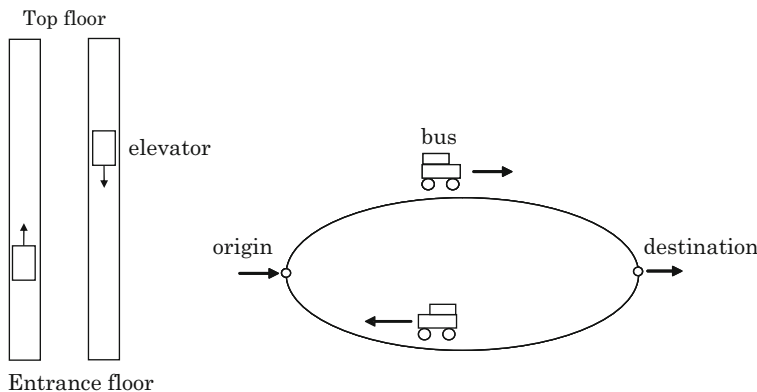
Model and Similarity

We consider the dynamic model of shuttle bus system which mimics the service of buses shuttling between the origin and the destination (e.g., an airport and a railway station). A bus carries the passengers from the origin to the destination. This shuttle bus model has the realistic background. We model the public transport system of shuttle buses as follows. Shuttle buses repeatedly go back and forth between the starting point (origin) and the destination. The starting point is the only position that passengers can take a bus. Passengers board a

bus at the origin, and then the bus starts from the origin. The bus moves toward the destination. When the bus arrives at the destination, all riding passengers leave the bus. After all passengers gets off the bus, the bus leaves the destination and returns to the origin. This process is repeated. This dynamics of the shuttle bus system can be described in terms of the nonlinear map models.

Also, we consider the case in which a bus serves repeatedly between two terminals A and B. Passengers come into each terminal independently. Passengers can board a bus at both terminals. Each terminal is an origin and a destination. When the tram or airplane arrives at terminals periodically, the passengers come into bus terminals periodically. The dynamics of the bus system is described in terms of the extended circle map.

One can consider the bus system that a few buses shuttle between the origin and destination according as a few buses pass freely over other buses. One compares this bus system with the elevator system that a few elevators shuttle between the entrance floor and the top floor. Passengers board the elevators at the entrance floor and passengers leave the elevators at the top floor. Then, the bus system has the similarity to elevators serving between the entrance floor and top floor. The bus dynamics is the same as that of elevators. Figure 1 shows the schematic illustration of two shuttle buses and two elevators (Nagatani 2002d, 2003b, 2004, 2013a, 2015, 2016a).



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 1 Schematic illustration for transport systems of two shuttle buses and two elevators. Two shuttle buses shuttle repeatedly

between the origin and the destination. A shuttle bus passes freely over another bus. Two elevators move from the ground floor to the top floor. There is a similarity between transport systems of buses and elevators

Also, one can consider the bus system that a few buses shuttle between the origin and destination and each bus never passes over other buses. One can compare this bus system with the tram system that a few trams make repeated roundtrips between the origin and destination. Each tram cannot pass over other trams. Passengers board the tram at the origin, the tram moves toward the destination, and passengers leave the tram at the destination. Therefore, the tram system has a similarity to the bus system. The motion of trams has the same dynamics as the bus system (Nagatani 2003c, 2007b). In the ferry boat system, the dynamic motion of ferry boats is similar to the bus transport system, and this dynamics can be described in terms of the nonlinear map model (Nagatani 2007c). In the air transport system, the dynamics of airplane is also described in the nonlinear map model (Nagatani 2013b).

One can consider the case that an operator controls the number of buses at the origin according to the number of passengers. We present the dynamic model of the bus system with increase or decrease of buses.

Here, we describe some dynamic models for the bus system since the elevator and tram systems are similar to the bus system.

Nonlinear Map Model

We consider a transport system of a single bus which carries passengers from the origin to the destination repeatedly. We describe the dynamic model of the single bus system by the nonlinear map. We assume that all the passengers waiting at the origin can board the bus. New passengers arrive at the origin with inflow rate μ [persons/minute]. The arrival time at the origin and trip n is defined by $t(n)$. So $\mu(t(n) - t(n-1))$ is the number of passengers that have arrived since the bus left the origin. The number of passengers boarding the bus at trip n is expressed by

$$B(n) = \mu(t(n) - t(n-1)). \quad (1)$$

The passengers boarding the next coming bus are those waiting at the origin until the arrival of

the bus. After the bus arrives at the origin, the passengers coming into the origin cannot board the bus. In this case, the passengers getting on the bus are given by Eq. 1. If one takes into account capacity $F_{\max}$ of a bus, Eq. 1 changes:

$$B(n) = \min[F_{\max}, \mu(t(n) - t(n-1))]. \quad (2)$$

We assume that the time which takes passengers to board the bus is proportional to the number of the passengers. The time is given by $\gamma B(n)$ where γ is the time which takes one person to board a bus. Generally, as the number of passengers increases, the time for passengers to board the bus gets longer. Similarly, the time which takes passengers to leave the bus is given by $\beta B(n)$ where β is the time which takes one person to leave a bus.

A bus driver can control the speed to retrieve the delay. A bus is accelerated or decelerated according to the tour time $\Delta t(n) (= t(n) - t(n-1))$. The speed depends on the tour time. The moving time of the bus is $2L/v(\Delta t)$ where L is the distance between two terminals (the origin and the destination) and $v(\Delta t)$ is the speed of bus. The tour time equals the sum of these periods. The speed is determined by the tour time $\Delta t(n)$ at trip n . Then, the arrival time $t(n+1)$ of the bus at the origin and trip $n+1$ is given by

$$t(n+1) = t(n) + (\gamma + \beta)\mu(t(n) - t(n-1)) + \frac{2L}{v(\Delta t(n))}. \quad (3)$$

The tour time $\Delta t(n+1)$ is given by

$$\Delta t(n+1) = (\gamma + \beta)\mu\Delta t(n) + \frac{2L}{v(\Delta t(n))}. \quad (4)$$

By dividing time by characteristic time $2L/v_0$ (v_0 : reference speed), one obtains the map in dimensionless form:

$$\Delta T(n+1) = \Gamma \Delta T(n) + \frac{1}{V(\Delta T(n))}, \quad (5)$$

where $\Delta T(n) = \Delta t(n)/(2L/v_0)$, $\Gamma = \mu(\gamma + \beta)$, and $V(\Delta T) = v(\Delta t)/v_0$.

Map (5) is a simple nonlinear map. The motion of the bus is determined by map (5). The dependence of speed on the tour time is given by the control method.

Piecewise Map Model

If a tour time is higher than the critical value, a bus speeds up abruptly and moves with high speed v_{high} . Otherwise, the bus moves with normal speed v_0 . Then, map (5) is given by

$$\Delta T(n+1) = \Gamma \Delta T(n) + 1 \quad \text{if } \Delta T(n) \leq T_c,$$

$$\Delta T(n+1) = \Gamma \Delta T(n) + \frac{1}{V_{\text{high}}} \quad \text{if } \Delta T(n) > T_c, \quad (6)$$

where $V_{\text{high}} = v_{\text{high}}/v_0$ and $T_c = t_c v_0/2L$.

Eq. 6 is the linear piecewise map. The characteristic of the map is determined by the fixed point (Nagatani 2002e, 2003d, e). If the map has a stable fixed point, the tour time approaches the value of fixed point after a sufficiently large time even if the initial value of the tour time takes any value. We study the fixed point of the map. In Fig. 2, we display the piecewise map (6) for some values of

dimensionless inflow rate Γ . Three curves are shown at $\Gamma = 0.1$, $\Gamma = 0.5$, and $\Gamma = 1.0$ where $v_{\text{high}}/v_0 = 2.0$ and $T_c = 1.5$. With increasing inflow rate Γ , the slope of the map increases. At $\Gamma = 0.1$, map (6) has a stable fixed point since the slope at the fixed point is less than one. At $\Gamma = 0.5$, map (6) does not have a stable fixed point but has an unstable fixed point because the absolute value of the slope at the fixed point is higher than one. At $\Gamma \geq 1$, map (6) has no fixed point. If the map has a stable fixed point, the tour time approaches the value of the stable fixed point after sufficiently large time. The shuttle bus moves with the constant value of the tour time. If the map has an unstable fixed point, the tour time does not approach the constant value of the fixed point after sufficiently large time, but it oscillates around the value of the fixed point. The shuttle bus moves periodically with trips. When the map does not have fixed points, the tour time diverges with increasing trip. The shuttle bus gets more and more late with increasing trip because the passengers increase with increasing tour time.

The fixed point is given by

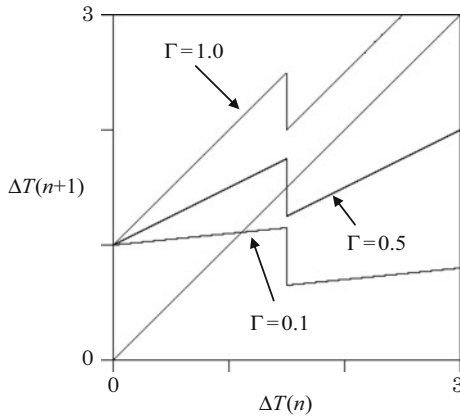
$$\begin{aligned} \text{stable fixed point } \Delta T_f &= \frac{1}{1-\Gamma} \quad \text{for } 0 \leq \Gamma < \frac{T_c-1}{T_c}, \\ \text{unstable fixed point } \Delta T_f &= T_c \quad \text{for } \frac{T_c-1}{T_c} \leq \Gamma < \frac{2T_c-1}{2T_c}, \\ \text{stable fixed point } \Delta T_f &= \frac{1}{2(1-\Gamma)} \quad \text{for } \frac{2T_c-1}{2T_c} \leq \Gamma < 1, \end{aligned} \quad (7)$$

where $v_{\text{high}}/v_0 = 2.0$.

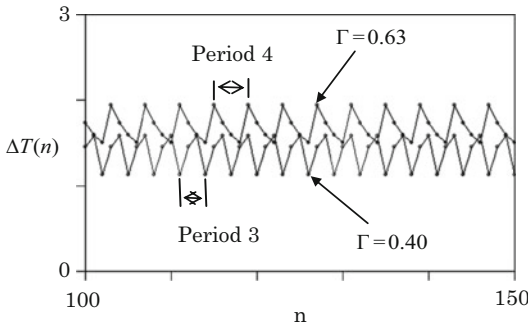
We show the typical behavior of the tour time with varying the inflow rate. Figure 3 shows the plots of tour time $\Delta T(n)$ against trip n at inflow rates $\Gamma = 0.40$ and $\Gamma = 0.63$ where $v_{\text{high}}/v_0 = 2.0$ and $T_c = 1.5$. The tour time oscillates with period 3 at $\Gamma = 0.40$ and period 4 at $\Gamma = 0.63$. Figure 4 shows the plot of the tour time against the inflow rate for $100 < n < 1000$. For $0 \leq \Gamma < 1/3$, the tour time displays a single curve, and its value is consistent

with the value of the stable fixed point. For $1/3 \leq \Gamma < 2/3$, the tour time takes multi-values. If the tour time takes triple values, it oscillates with period 3. For $2/3 \leq \Gamma < 1$, the tour time displays a single curve, and its value agrees with the value of the stable fixed point.

Generally, the boarding and leaving times are proportional to the power of the tour time. The linear piecewise map (6) is extended to the non-linear piecewise map:



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 2 Three curves of piecewise map (6) at $\Gamma = 0.1$, $\Gamma = 0.5$, and $\Gamma = 1.0$. With increasing inflow rate Γ , the slope of the map increases. At $\Gamma = 0.1$, map (6) has a stable fixed point. At $\Gamma = 0.5$, map (6) has an unstable fixed point. At $\Gamma \geq 1$, map (6) has no fixed point

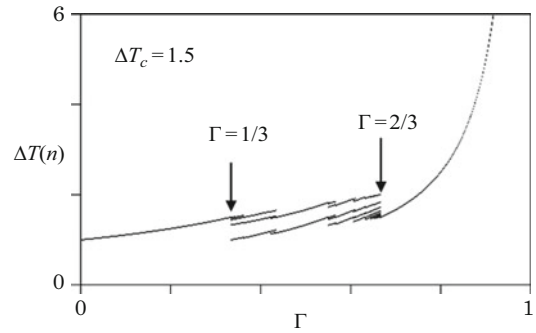


Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 3 Plots of tour time $\Delta T(n)$ against trip n at inflow rates $\Gamma = 0.40$ and $\Gamma = 0.63$. The tour time oscillates with period 3 at $\Gamma = 0.40$ and period 4 at $\Gamma = 0.63$

$$\Delta T(n+1) = \Gamma \Delta T(n)^\alpha + 1 \quad \text{if } \Delta T(n) \leq T_c,$$

$$\Delta T(n+1) = \Gamma \Delta T(n)^\alpha + \frac{1}{V_{\text{high}}} \quad \text{if } \Delta T(n) > T_c, \quad (8)$$

where α is a positive real number. In the extended version, the tour time displays more



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 4 Plot of the tour time against inflow rate Γ for $100 < n < 1000$. For $0 \leq \Gamma < 1/3$ and $2/3 \leq \Gamma < 1$, the tour time displays a single curve, and its value is consistent with the value of the stable fixed point. For $1/3 \leq \Gamma < 2/3$, the tour time takes multi-values and displays the oscillation

complex behavior than map (6). The bus behaves not only periodically but also chaotically (Nagatani 2002e, 2003d). Figure 5 shows the plot of tour time $\Delta T(n)$ against inflow rate Γ at $\alpha = 2.0$. In the region between points a and b , the tour time oscillates periodically. In the region higher than point b , the tour time displays chaotic or quasiperiodic behaviors.

The bus driver can control the speed of the bus by other methods. For example, the driver speeds up more and more when tour time Δt approaches critical value t_c and the bus speed reaches maximum speed v_{max} after a while. We assume that the speed is controlled by the hyperbolic tangent function. In this case, map (6) changes to the hyperbolic tangent map.

Delayed Map Model

If the speed is controlled with delay time τ , the speed is described by $v(\Delta t(n - \tau))$. When $\tau = 1$, the speed is controlled by tour time $\Delta t(n - 1)$ at trip $n - 1$. The control of the speed is done per one round late. When $\tau = 2$, the speed is controlled by tour time $\Delta t(n - 2)$ at trip $n - 2$. In this case, the control of the speed is done per two rounds late. Then, the tour time for the bus system with delay τ is given by

$$\Delta T(n+1) = \Gamma \Delta T(n) + \frac{1}{V(\Delta T(n-\tau))}, \quad (9)$$

$$V(\Delta T(n-\tau)) = 1 + \frac{(v_{\max}/v_0 - 1)}{2} \{1.0 + \tanh[A(\Delta T(n-\tau) - T_c)]\}, \quad (10)$$

Eq. 9 represents a kind of delayed map. Thus, we obtain the delayed map model for the tour time in the shuttle bus transportation (Nagatani 2016b). The dynamics of a bus is described by delayed map (9). Delayed map (9) is a $(1 + \tau)$ -dimensional map. One-dimensional map with no delay changes to $(1 + \tau)$ -dimensional map (9) by introducing the delay of the speed control. The characteristics of the high-dimensional map are very complex, and it is difficult to derive the fixed point analytically. The delayed map has not been studied, but the delayed logistic map has been studied by Masoller and Rosso (2011).

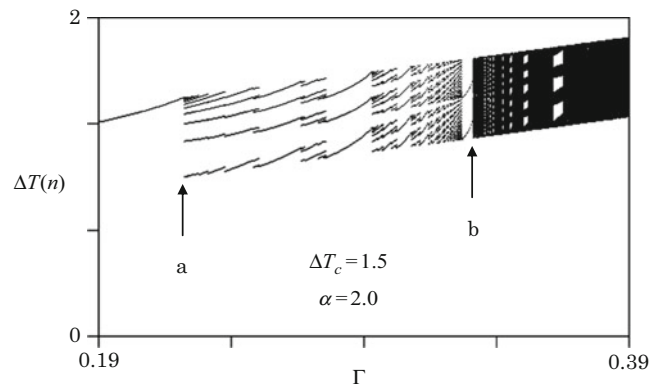
The effect of delay τ on the arrival time and bus schedule is studied. For comparison, the tour time is shown in the case of no delay $\tau = 0$. Figure 6 shows the plots of tour time $\Delta T(n)$ against inflow rate Γ at $A = 5$ and $A = 8$ and $\tau = 0$ from $n = 100$ to $n = 1000$ where $v_{\max}/v_0 = 2.0$ and $T_c = 1.5$. The tour time at $A = 5$ converges a single value in all the regions and is indicated by a dashed line. At $A = 8$, the tour time converges a single value except for the region between points a and b and displays a single curve except for the region between points a and b . The tour time agrees with the fixed point except for the region between points

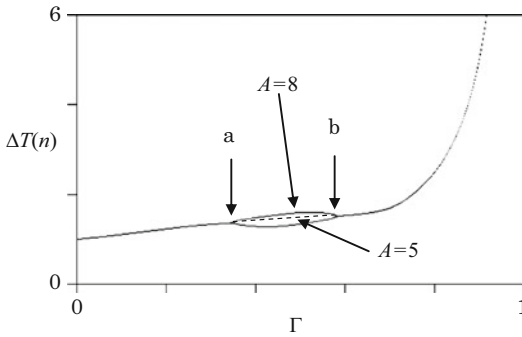
a and b . However, the tour time does not converge a single value in the region between points a and b but has double values. This is due to the unstable fixed point in the region between points a and b . The tour time oscillates periodically with period 2.

Figure 7 shows the plot of tour time $\Delta T(n)$ against inflow rate Γ at $A = 8$ and $\tau = 1$ from $n = 100$ to $n = 1000$ where $v_{\max}/v_0 = 2.0$ and $T_c = 1.5$. The tour time converges a single value except for the region between points a and b and displays a single curve except for the region between points a and b . However, the tour time does not converge a single value in the region between points a and b but has many values. In the region between points c and d , the tour time takes four values and displays a periodic motion of period 4 with varying trip n . By introducing delay $\tau = 1$, the tour time profile (Fig. 6) in the non-delayed map ($\tau = 0$) changes to that in Fig. 7.

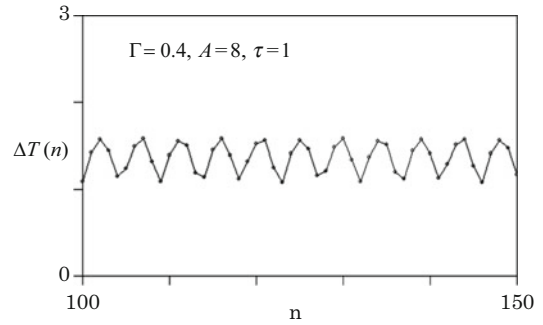
Figure 8 shows the plot of tour time $\Delta T(n)$ against trip n from $n = 100$ to $n = 150$ at $A = 8$, $\tau = 1$ and $\Gamma = 0.4$ where $v_{\max}/v_0 = 2.0$ and $T_c = 1.5$. The tour time exhibits a roughly periodic motion with period 9 and varies irregularly in detail. Figure 9 shows the plot of tour time $\Delta T(n+1)$ against $\Delta T(n)$ from $n = 100$ to $n = 1000$ at $A = 8$, $\tau = 1$ and $\Gamma = 0.4$ where $v_{\max}/v_0 = 2.0$ and $T_c = 1.5$. The return map in Fig. 9 corresponds to Fig. 8. The return map shows a smooth closed curve. The smooth closed curve is similar to that observed in the circle map (He et al. 1985). The smooth closed curve is observed in the quasiperiodic motion. The dynamic

Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 5 Plot of tour time $\Delta T(n)$ against inflow rate Γ at $\alpha = 2.0$. In the region between points a and b , the tour time oscillates periodically. In the region higher than point b , the tour time displays chaotic or quasiperiodic behaviors

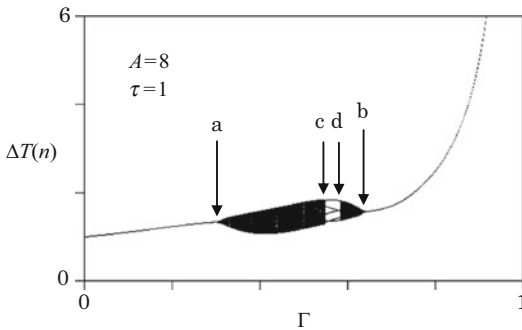




Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 6 Plots of tour time $\Delta T(n)$ against inflow rate Γ at $A = 8$ and $A = 5$ where $\tau = 0$. The tour time at $A = 8$ converges a single value except for the region between points a and b . In the region between points a and b , the tour time oscillates periodically with period 2. The tour time at $A = 5$ converges a single value in all the regions and is indicated by a *dashed line*



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 8 Plot of tour time $\Delta T(n)$ against trip n from $n = 50$ to $n = 150$ at $A = 8$, $\tau = 1$ and $\Gamma = 0.4$. The tour time exhibits a roughly periodic motion with period 9 and varies irregularly in detail

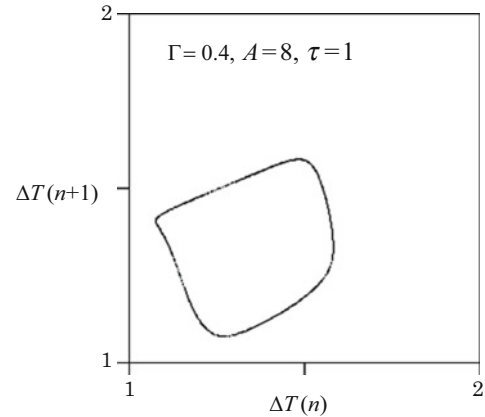


Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 7 Plot of tour time $\Delta T(n)$ against inflow rate Γ at $A = 8$ and $\tau = 1$. The tour time displays a *single curve* except for the region between points a and b . The tour time does not converges a single value in the region between points a and b but has many values. In the region between points c and d , the tour time takes four values and displays a periodic motion of period 4 with varying trip n . By introducing delay $\tau = 1$, the tour time profile (Fig. 6) in the non-delayed map ($\tau = 0$) changes to that in Fig. 7

motion of the bus displays the quasiperiodicity by introducing the speedup delay. Thus, the speedup delay has the important effect on the bus motion.

Extended Circle Map Model

Generally, the arrival rate of the passengers is periodic or irregular. When a bus route is



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 9 The return map of the tour time in Fig. 8. The return map draws a *smooth closed curve*

connected with other routes, the arrival rate of the passengers is periodic at the transfer point. We consider the bus system that the arrival rate of passengers is periodic at the terminals. In realistic situations, the demand (the arrival rate of the passengers) is not simply periodic but quasiperiodic or irregular. If the fluctuation of the demand is weak around the periodicity, the quasiperiodic demand can be taken into account

as the superposition of the periodic part and noise part. The periodic passenger arrival rate simulation helps us understand the bus behavior for more realistic situations.

A single bus serves repeatedly between two terminals A and B. Passengers come into each terminal periodically. Passengers coming into terminal A are independent on those coming into terminal B. Passengers board the bus at terminal A, the bus moves toward terminal B, and all passengers get off the bus when it arrives at terminal B. As soon as the bus is empty, passengers at terminal B board the bus, the bus moves toward terminal A, and all passengers get off the bus when it arrives at terminal A. The process is repeated. In the previous model, the starting terminal is the only way to board the bus. Here we extend the previous model to take into account boarding at terminal B. The bus shuttles repeatedly between two terminals. When the tram or airplane arrives at terminals periodically, the passengers come into terminals periodically. Then, passengers board on the bus at terminals with periodic inflows.

The dynamic model for the shuttle bus with the periodic inflows is described by the non-linear map (Nagatani and Naito 2011, 2012; Nagatani 2013c, 2016c). Passengers come into terminal A (B) at the inflow rate with period $t_{s,A}$ ($t_{s,B}$). Define the number of passengers boarding the bus at terminal A(B) on trip n by $B_A(n)$ ($B_B(n)$). So $\gamma B_{A(B)}(n)$ is the amount of time needed to board all the passengers at terminal A(B). The moving time of the bus between terminals A and B is L/V where L is the length between terminals A and B and V is the mean speed of bus. The stopping time at terminal A(B) to leave off the passengers is $\beta B_{A(B)}(n)$ where parameter β is the time which takes one passenger to leave the bus. The tour time equals the sum of these periods. When all passengers boarding the bus at terminal A get off the bus at terminal B, the bus becomes empty. As soon as the bus is empty, new passengers are ready to board the bus. At once, new passengers waiting at terminal B begin to board the bus. We refer

the time as the ready time. Therefore, the ready time $t_B(n)$ of the bus at terminal B on trip n is given by

$$t_B(n) = t_A(n) + \gamma B_A(n) + \frac{L}{V} + \beta B_A(n). \quad (11)$$

The ready time $t_A(n+1)$ of the bus at terminal A on trip $n+1$ is given by

$$\begin{aligned} t_A(n+1) = t_A(n) + \gamma B_A(n) + \frac{L}{V} \\ + \beta B_A(n) + \gamma B_B(n) + \frac{L}{V} \\ + \beta B_B(n). \end{aligned} \quad (12)$$

Define $W_{A(B)}(n)$ as the number of passengers waiting at terminal A(B) just before the bus arrives at terminal A(B) on trip n . New passengers arrive at terminal A(B) at rate $\mu_{A(B)}(t)$. So $W_{A(B)}(n)$ is the number of passengers that have arrived since the previous bus left terminal A(B). $W_A(n)$ and $W_B(n)$ are expressed by

$$W_A(n) = \int_{t_A(n-1)}^{t_A(n)} \mu_A(t) dt, \quad (13)$$

$$W_B(n) = \int_{t_B(n-1)}^{t_B(n)} \mu_B(t) dt. \quad (14)$$

The periodic inflow rate is expressed by

$$\begin{aligned} \mu_{A(B)}(t) = \mu_{0,A(B)} \\ \left\{ 1 + b_{A(B)} \cos \left(\frac{2\pi \left(t + \varphi_{A(B)} \right)}{t_{s,A(B)}} \right) \right\}, \end{aligned} \quad (15)$$

where $t_{s,A(B)}$ is the period of the inflow rate at terminal A(B), $\phi_{A(B)}$ is the phase shift of the inflow rate at terminal A(B), and $b_{A(B)}$ is the amplitude of fluctuating inflow rate at terminal A(B). Since $\mu_{A(B)}(t) \geq 0, 0 \leq b_{A(B)} \leq 1$. By

substituting Eqs. 13 and 14 into Eqs. 11 and 12 and dividing time by the characteristic time $2L/V$, one obtains the nonlinear maps of dimensionless ready time:

$$\begin{aligned} T_A(n+1) = & T_A(n) + \Gamma_{0,A}(T_A(n) - T_A(n-1)) + \Gamma_{0,B}(T_B(n) - T_B(n-1)) \\ & + \frac{\Gamma_{0,A}b_AT_{s,A}}{2\pi} \left\{ \sin\left(\frac{2\pi(T_A(n) + \Phi_A)}{T_{s,A}}\right) - \sin\left(\frac{2\pi(T_A(n-1) + \Phi_A)}{T_{s,A}}\right) \right\}, \\ & + \frac{\Gamma_{0,B}b_BT_{s,B}}{2\pi} \left\{ \sin\left(\frac{2\pi(T_B(n) + \Phi_B)}{T_{s,B}}\right) - \sin\left(\frac{2\pi(T_B(n-1) + \Phi_B)}{T_{s,B}}\right) \right\} + 1 \end{aligned} \quad (16)$$

$$\begin{aligned} T_B(n) = & T_A(n) + \Gamma_{0,A}(T_A(n) - T_A(n-1)) \\ & + \frac{\Gamma_{0,A}b_AT_{s,A}}{2\pi} \left\{ \sin\left(\frac{2\pi(T_A(n) + \Phi_A)}{T_{s,A}}\right) - \sin\left(\frac{2\pi(T_A(n-1) + \Phi_A)}{T_{s,A}}\right) \right\} + 0.5, \end{aligned} \quad (17)$$

where $T_{A(B)}(n) \equiv t_{A(B)}(n)V/2L$, $T_{s,A(B)} \equiv t_{s,A(B)}V/2L$, and $\Gamma_{0,A(B)} \equiv (\gamma + \beta)\mu_{0,A(B)}$.

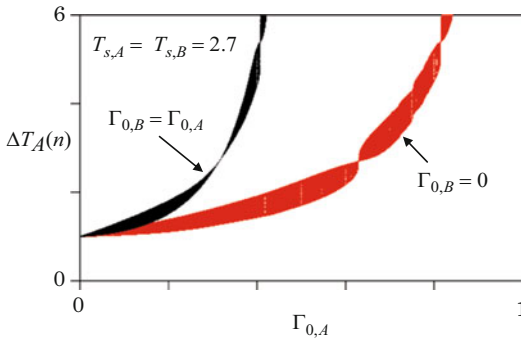
Nonlinear maps (16) and (17) of the ready times are the extended version of the circle map (He et al. 1985). Thus, the dynamics of the shuttle bus is described in terms of the extended circle maps (16) and (17) for the shuttle bus with two periodic inflow rates. The dynamic property of the map is controlled by six parameters (loading parameters $\Gamma_{0,A}$ and $\Gamma_{0,B}$, periods $T_{s,A}$ and $T_{s,B}$, and phases Φ_A and Φ_B), when $b_A = b_B = 1$.

When $\Gamma_{0,B} = 0$, there are no passengers boarding the bus at terminal B; terminal A is the origin, terminal B is the destination, and passengers board only at terminal A and get off only at terminal B.

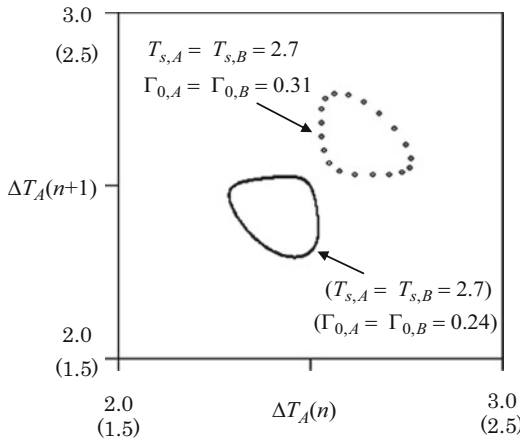
In the case of $T_{s,A} = T_{s,B}$ and $\Phi_A = \Phi_B = 0$, the inflow of passengers at terminal B synchronizes with that at terminal A. Both inflows of passengers have the same period and phase shift. Figure 10 shows the plot of tour time $\Delta T_A(n)$ versus loading parameter $\Gamma_{0,A}$ for arrival numbers from $n = 1000$ to $n = 2000$ for inflow periods $T_{s,A} = T_{s,B} = 2.7$. Black dots indicate the tour time for $\Gamma_{0,A} = \Gamma_{0,B}$. For comparison, we calculate the tour time in the case that passengers do not come into terminal B but come only into terminal A. Red dots indicate the tour time for

$\Gamma_{0,B} = 0$. The tour time varies in a complex manner with the loading parameter. The tour time takes on multiple values except for a few points. By adding the inflow at terminal B to the bus transport, the tour time becomes about two times. The tour time increases with the loading parameter and diverges at $\Gamma_{0,A} = 0.5$. The value at the divergence point is half for the tour time of $\Gamma_{0,B} = 0$. The tour time drastically changes by adding the inflow at terminal B. The return map is shown for the shuttle bus transport. Figure 11 shows the plots of $\Delta T_A(n+1)$ versus $\Delta T_A(n)$ for $T_{s,A} = T_{s,B} = 2.0$, $\Gamma_{0,A} = \Gamma_{0,B} = 0.31$, $T_{s,A} = T_{s,B} = 2.7$, and $\Gamma_{0,A} = \Gamma_{0,B} = 0.24$. Return map on the upper side in Fig. 11 displays 19 points because the tour time varies periodically with period 19. Return map on the bottom of Fig. 11 shows the smooth closed curve. Therefore, the tour time varies quasiperiodically. Furthermore, the return map is calculated for various values of both loading parameter and inflow period. The return map displays either discrete points or smooth closed curve. The chaotic motion does not occur. Thus, the shuttle bus shows either periodic motion or quasiperiodic motion in the case of $T_{s,A} = T_{s,B}$ and $\Phi_A = \Phi_B = 0$.

In the case of $T_{s,A} \neq T_{s,B}$ and $\Phi_A = \Phi_B = 0$, passengers come periodically into terminal A with

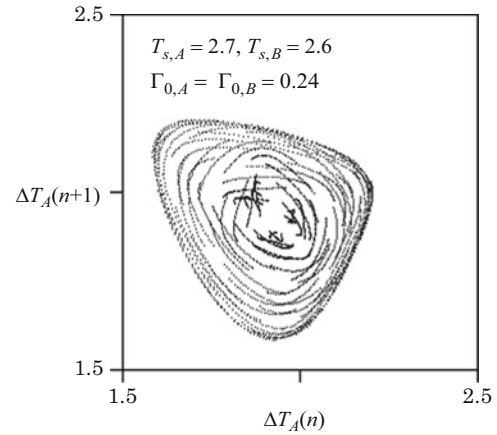


Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 10 Plot of tour time $\Delta T_A(n)$ versus loading parameter $\Gamma_{0,A}$ for inflow periods $T_{s,A} = T_{s,B} = 2.7$. For comparison, the tour time is shown for $\Gamma_{0,B} = 0$. The tour time takes on multiple values except for a few points



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 11 Plot of $\Delta T_A(n+1)$ versus $\Delta T_A(n)$ at $T_{s,A} = T_{s,B} = 2.0$ and $\Gamma_{0,A} = \Gamma_{0,B} = 0.31$. The return map displays 19 points because the tour time varies periodically with period 19. Plot of $\Delta T_A(n+1)$ versus $\Delta T_A(n)$ for $T_{s,A} = T_{s,B} = 2.7$ and $\Gamma_{0,A} = \Gamma_{0,B} = 0.24$. The return map shows the *smooth closed curve*. The tour time varies quasiperiodically

the period different from that of the inflow of passengers into terminal B and the phase shift of inflow A is the same as that of inflow B. Generally, the period of inflow at terminal A is different from that at terminal B because terminal A is



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 12 Return map at $T_{s,A} = 2.7$, $T_{s,B} = 2.6$, and $\Gamma_{0,A} = \Gamma_{0,B} = 0.24$. By changing period $T_{s,B}$ from $T_{s,B} = 2.7$ to $T_{s,B} = 2.6$, the return map changes from the *smooth closed curve* in Fig. 11 to the breakup of the torus in Fig. 12. The quasiperiodicity in Fig. 11 changes to a chaotic motion

connected with the bus route different from terminal B. The effect of the period difference on the bus motion is studied under the condition of the same phase shift. Figure 12 is the return map for $T_{s,A} = 2.7$, $T_{s,B} = 2.6$, and $\Gamma_{0,A} = \Gamma_{0,B} = 0.24$. By changing period $T_{s,B}$ from $T_{s,B} = 2.7$ to $T_{s,B} = 2.6$, the return map changes from the smooth closed curve in Fig. 11 to the breakup of the torus in Fig. 12. The quasiperiodicity in Fig. 11 changes to a chaotic motion. Thus, the quasiperiodicity of the bus changes to the chaotic motion by varying period $T_{s,B}$. The period difference induces the chaotic motion of the bus.

Combined Map Model

When passengers come periodically at the origin, and the bus driver controls the speed to retrieve the delay, the dynamics is described by the extended circle map combined with the hyperbolic tangent map. The nonlinear map is called the combined map (Nagatani 2013d, 2017). The combined map of the dimensionless arrival time is given by

$$\begin{aligned}
T(n+1) = & T(n) + \Gamma_0(T(n) - T(n-1)) \\
& + \frac{\Gamma_0 b T_s}{2\pi} \left\{ \sin\left(\frac{2\pi(T(n) + \Phi)}{T_s}\right) - \sin\left(\frac{2\pi(T(n-1) + \Phi)}{T_s}\right) \right\} \\
& + \frac{1}{1 + \frac{1}{2} \left(\frac{v_{\max}}{v_0} - 1 \right) \{1 + \tanh[A(\Delta T(n) - T_c)]\}}
\end{aligned} \tag{18}$$

Map (18) consists of both map (9) and map (16) with $\Gamma_{0,B} = 0$. The shuttle bus shows the periodic, quasiperiodic, and chaotic motions by varying both periodic inflow and speed control. The chaotic motion does not occur only in the hyperbolic map or in the extended circle map, but the chaotic motion occurs by combining the hyperbolic tangent map with the extended circle map. Also, one is able to introduce the delay into the combined map. Then, combined map (18) is more complex delayed combined map.

Freely Passing Buses

We consider the case that a bus passes other buses freely when the bus reaches those. The dynamic model is presented for the fluctuation of riding passengers in a few shuttle buses (Nagatani 2003f, g, h, 2006b). Now we are looking at the case that passengers are only boarding at the origin and alighting at the destination. When the time headway between buses is shorter (longer), the awaiting passengers at the origin are fewer (more). The passengers boarding at the origin are few (many). In the result, the boarding and getting off times are shorter (longer). The bus may pass over the previous bus. Thus, buses run a neck-and-neck race with other buses. If the awaiting passengers are superior to the maximum capacity of the bus, the passengers corresponding to the capacity board the bus, and the remaining passengers wait for the next coming bus. Bus i moves with speed v_i . Then, the dimensionless arrival time of bus i at the origin is given by

$$T_i(m+1) = T_i(m) + \Gamma_0 B_i(m) + \frac{v_0}{v_i}, \tag{19}$$

with

$$\begin{aligned}
W_i(m) = & W_{i'}(m') - B_{i'}(m') \\
& + \Pi(T_i(m) - T_{i'}(m'))
\end{aligned} \tag{20}$$

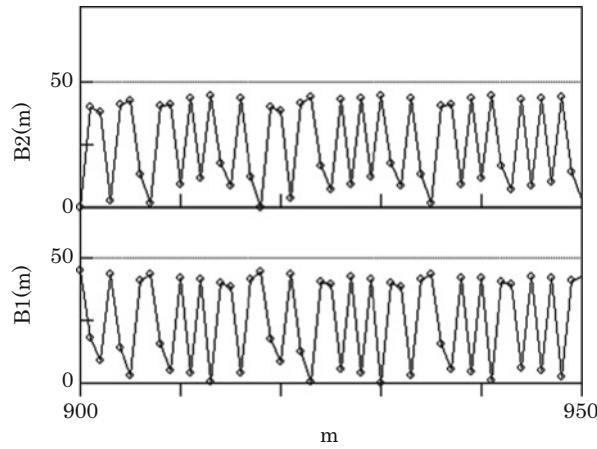
and

$$B_i(m) = \min[F_{i,\max}, W_i(m)], \tag{21}$$

where $T_i(m) \equiv \frac{t_i(m)v_0}{2L}$, $\Gamma_0 \equiv \frac{(\gamma+\eta)v_0}{2L}$, and $\Pi \equiv \frac{\mu 2L}{v_0}$. Subscript i' indicates the next coming bus after bus i .

The dynamics of the buses is described by map (19) with (20) and (21). The map is iterated simultaneously for M buses. The dynamical property of the map is controlled by four parameters: loading parameter $\Gamma(=\Gamma_0\Pi)$, capacity $F_{i,\max}$, bus speed v_i/v_0 , and bus number M . When perspective passengers increase, the value of loading parameter becomes high. For a few buses, the order of buses changes with trip m , and the time headway between buses changes from trip to trip because a bus passes other buses or is outstripped by other buses. As a result, the riding passengers also change from bus to bus. The buses exhibit a complex behavior.

The number of riding passengers with varying trips is calculated by iterating map (19) for the typical case of two buses. Figure 13 shows the plots of the numbers $B1(m)$ and $B2(m)$ of riding passengers on buses 1 and 2 against trip m from trip $m = 900$ to $m = 950$ for $\Pi = 40$ where $M = 2$, $\Gamma_0 = 0.01$, $v_1 = v_2 = v_0$, and $F_{1,\max} = F_{2,\max} = 50$. The dotted line indicates the



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 13 Plots of the numbers $B1(m)$ and $B2(m)$ of riding passengers on buses 1 and 2 against trip m from trip $m = 900$ to

$m = 950$ for $\Pi = 40$. The dotted line indicates the maximum capacity $F_{1(or2), \max}$. The numbers $B1(m)$ and $B2(m)$ of riding passengers are lower than $F_{1(or2), \max}$. The numbers of riding passengers vary irregularly

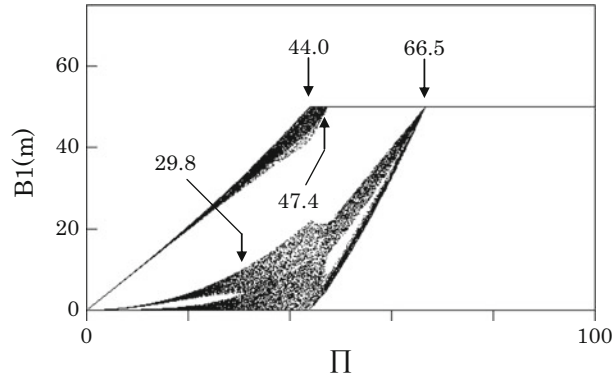
maximum capacity $F_{1(or2), \max}$. The numbers $B1(m)$ and $B2(m)$ of riding passengers cannot be superior to $F_{1(or2), \max}$. For a low value of loading parameter, the numbers of riding passengers change irregularly from zero to a value lower than the maximum capacity $F_{1(or2), \max}$. For a high value of loading parameter, the numbers of riding passengers change irregularly from non-zero to the maximum capacity $F_{1(or2), \max}$. The buses carry frequently the full load of passengers.

The number of riding passengers changes by synchronizing the time headway. The time headway changes alternately from low to high values. Both time headway and riding passengers vary almost periodically. However, these quantities exhibit an irregularity. When two buses arrive simultaneously at the origin, the time headway becomes zero. In the result, the riding passengers become zero because the waiting passengers are proportional to the time headway for no queuing at the origin.

Figure 14 shows the plot of the number $B1(m)$ of riding passengers on bus 1 against loading parameter Π from sufficiently large trip $m = 900$ to $m = 1000$ without noises where $M = 2$, $\Gamma = 0.01$, and $F_{1, \max} = F_{2, \max} = 50$. For low values of loading parameter Π , both number of

riding passengers and time headway take the three values, and their distributions exhibit the three peaks around the values. With increasing loading parameter Π , the localized distributions extend around the peaks and become the two extended distributions for $29.8 < \Pi < 47.4$. When loading parameter Π is higher than 47.4, the two extended distribution breaks down three distributions again. If loading parameter Π is higher than 66.5, the time headway keeps a constant value. For $\Pi < 66.5$, the time headway exhibits the chaotic motion. The distribution of the time headway exhibits the similar behavior to the number of riding passengers. When the loading parameter is higher than $\Pi = 44.0$, the number of riding passengers saturates some times at $F_{1(or2), \max}$. For $47.4 < \Pi < 66.5$, either bus 1 or bus 2 carries the full load of passengers. When loading parameter Π is higher than 66.5, both numbers $B1(m)$ and $B2(m)$ saturate at the maximum capacity. At $\Pi_c = 66.5$, two buses become the full load. Therefore, the dynamical transition from the chaotic to regular motions at $\Pi_c = 66.5$ is induced by the full-load buses.

In the case of freely passing buses, the behavior of buses exhibits a deterministic chaos even if there are no noises. The chaotic motion of buses



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 14 Plot of the number $B1(m)$ of riding passengers on bus 1 against loading parameter Π from sufficiently large trip $m = 900$ to $m = 1000$ without noises. For low values of loading parameter Π , the number of riding passengers takes the three values, and their distributions exhibit the three peaks around the values. With increasing loading parameter Π , the localized distributions extend around the peaks and become the two extended distributions for $29.8 < \Pi < 47.4$. When loading parameter Π is higher than 47.4, the two extended distribution breaks down three

distributions again. If loading parameter Π is higher than 66.5, the time headway keeps a constant value. For $\Pi < 66.5$, the time headway exhibits the chaotic motion. When the loading parameter is higher than $\Pi = 44.0$, the number of riding passengers saturates some times at $F_{1-(or2), \max}$. For $47.4 < \Pi < 66.5$, either bus 1 or bus 2 carries the full load of passengers. When loading parameter Π is higher than 66.5, both numbers $B1(m)$ and $B2(m)$ saturate at the maximum capacity. At $\Pi_c = 66.5$, two buses becomes the full load. Therefore, the dynamical transition from the chaotic to regular motions at $\Pi_c = 66.5$ is induced by the full load buses

is induced by the neck-and-neck race between buses. The dynamic behavior of buses is definitely different from that of the single bus.

$$T_1(n+1) = T_1(n) + \Gamma \{T_1(n) - T_N(n-1)\} + 1 \quad \text{if } T_1(n+1) \geq T_N(n). \quad (22)$$

If bus 1 catches bus N , bus 1 restarts after delay time $T_{\min}$:

$$T_1(n+1) = T_N(n) + T_{\min} \quad \text{if } T_1(n+1) < T_N(n). \quad (23)$$

Similarly, the arrival time $T_i(n+1)$ of bus i ($2 \leq i \leq N$) at starting point and trip $n+1$ is given by

$$T_i(n+1) = T_i(n) + \Gamma \{T_i(n) - T_{i-1}(n-1)\} + 1 \quad \text{if } T_i(n+1) \geq T_{i-1}(n). \quad (24)$$

$$T_i(n+1) = T_{i-1}(n) + T_{\min} \quad \text{if } T_i(n+1) < T_{i-1}(n). \quad (25)$$

We consider the case that N shuttle buses do not pass each other on the route. The dynamical model of N shuttle buses is presented (Nagatani 2003i, 2005b). All buses are numbered from 1 (leading bus) to N (latest bus) in regular order. The buses start in regular order of 1, 2, 3, ..., N at the starting point after perspective passengers board buses, they move toward the destination, all currently boarding passengers leave the bus after the bus arrives at the destination, and the buses return to the starting point. If a shuttle bus catches the other bus, it stops and restarts after delay time $T_{\min}$ in order to keep the appropriate time headway. The bus does not change the speed except for stopping. The bus keeps constant speed until it stops. The new passengers arrive at the starting point at rate μ . ΔT is the time headway between itself and its predecessor. When bus 1 does not catch bus N , the arrival time $T_1(n+1)$ of bus 1 at starting point at trip $n+1$ is given by

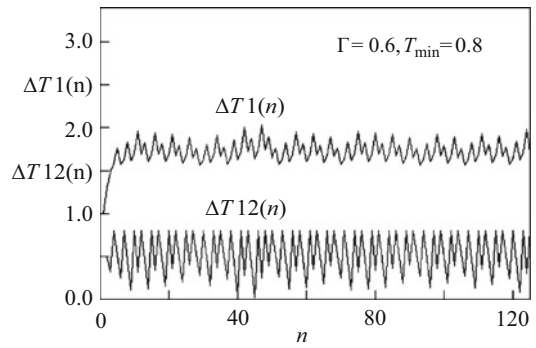
The bus motion depends on loading parameter Γ , delay time $T_{\min}$, and number N of buses. With increasing loading parameter Γ , the bus slows down since more passengers board the bus. With

decreasing delay time $T_{\min}$, the following bus becomes faster since less passengers board the bus. N shuttle buses exhibit a complex behavior and dynamical transitions by varying values of the three parameters.

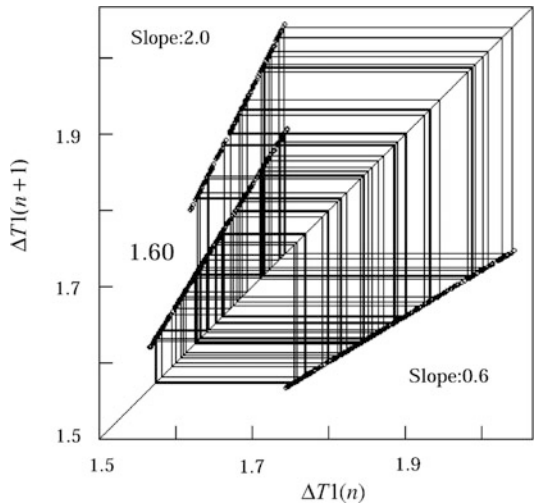
The behavior of two buses is derived by the use of iterates of Eqs. 22, 23, 24, and 25. Figure 15 shows the time evolutions of the tour time $\Delta T_1(n)$ and the time headway $\Delta T_{12}(n)$ for delay time $T_{\min} = 0.8$ under the constant value $\Gamma = 0.6$ of loading parameter where $\Delta T_1(n) = T_1(n) - T_1(n-1)$ and $\Delta T_{12}(n) = T_2(n) - T_1(n)$. For $T_{\min} = 0.8$, two buses oscillate irregularly. The oscillation exhibits a chaotic behavior. Figure 16 shows the plot of tour time $\Delta T_1(n+1)$ against $\Delta T_1(n)$ over $n = 100-1000$. This is the return map for Fig. 15. A circle indicates a value of tour time $\Delta T_1(n)$. The solid lines represent the trajectories of time series plot over $n = 100-300$. This shows that the bus behavior is a chaotic motion. The chaotic values of tour time lie on three straight lines with slopes 0.60, 1.60, and 2.0.

The dependence of tour time $\Delta T_1(n)$ on delay time $T_{\min}$ is derived to study the dynamical transitions to the complex motions. Figure 17 shows the plot of the values of tour time $\Delta T_1(n)$ against delay time $T_{\min}$ for loading parameter $\Gamma = 0.6$. For a small value of $T_{\min}$, two buses move with a constant value of tour time, and bus 2 follows bus 1 with constant value $T_{\min}$ of time headway. When delay time $T_{\min}$ is less than the critical value $T_{\min} = 0.48$, the tour time converges to the constant value. If delay time $T_{\min}$ is higher than the critical value $T_{\min} = 0.48$, the bus oscillates periodically. The dynamical transition from the regular motion to the periodic motion occurs at $T_{\min} = 0.48$. At $T_{\min} = 0.71$, the distinct transition from double periodic motion to the chaotic motion occurs. The value of Lyapunov exponent changes from a negative value to a positive value at $T_{\min} = 0.71$.

The phase diagram (region map) is derived for the distinct dynamical states and the dynamical transitions. Figure 18 shows the phase diagram in phase space $(T_{\min}, \Gamma)$. The regular motion in which the tour time converges to a constant value is indicated by gray color. The periodic motion is represented by white color. The chaotic motion is indicated by the circle. With increasing



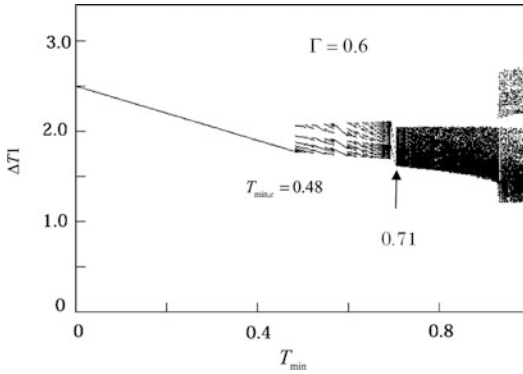
Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 15 Time evolutions of tour time $\Delta T_1(n)$ and time headway $\Delta T_{12}(n)$ for delay time $T_{\min} = 0.8$ under loading parameter $\Gamma = 0.6$. Two buses oscillate irregularly



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 16 The return map for Fig. 15. Plot of tour time $\Delta T_1(n+1)$ against $\Delta T_1(n)$. A circle indicates a value of tour time $\Delta T_1(n)$. The solid lines represent the trajectories of time series plot over $n = 100-300$

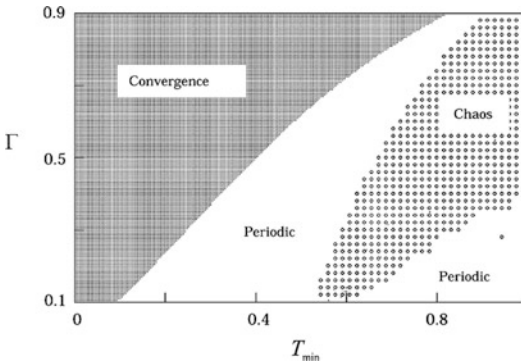
loading parameter Γ , the region of regular motion enlarges and the region of periodic motion shrinks. The complex behavior of two shuttle buses depends highly on both loading parameter Γ and delay time $T_{\min}$.

In the case of three shuttle buses, the behavior similar to two buses is observed. The shuttle bus system with no passing is definitely different from



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 17

Plot of tour time $\Delta T_1(n)$ against delay time $T_{\min}$ at loading parameter $\Gamma = 0.6$. For a small value of $T_{\min}$, two buses move with a constant tour time, and bus 2 follows bus 1 with constant value $T_{\min}$ of time headway. When delay time $T_{\min}$ is less than the critical value $T_{\min} = 0.48$, the tour time converges to the constant value. If delay time $T_{\min}$ is higher than the critical value $T_{\min} = 0.48$, the bus oscillates periodically. The dynamical transition from the regular motion to the periodic motion occurs at $T_{\min} = 0.48$. At $T_{\min} = 0.71$, the distinct transition from double periodic motion to the chaotic motion occurs



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 18

Phase diagram in phase space $(T_{\min}, \Gamma)$. The regular motion is indicated by gray color. The periodic motion is represented by white color. The chaotic motion is indicated by the circle. With increasing loading parameter Γ , the region of regular motion enlarges and the region of periodic motion shrinks. The complex behavior of two shuttle buses depends highly on both loading parameter Γ and delay time $T_{\min}$

that with passing freely. The passing has the important effect on the dynamic motion, traffic states, and dynamic transitions.

Increase or Decrease of Buses

We consider the case that an operator controls the number of buses at the origin according to the number of passengers (Nagatani 2011a, b, 2012, 2013e). $1 \sim N$ buses shuttle repeatedly between the starting point (origin) and the destination. Passengers arrive continuously at the origin with a rate. The number of buses increases one by one when the number of passengers is superior to the capacity, while the number of buses decreases one by one if the number of passengers is less than the capacity. For example, buses 1 and 2 start at the same time from the origin and move together. Buses 3- N stop at the origin and wait for a chance. If the number of passengers is higher than the total capacity, bus 3 runs with buses 1 and 2. Buses 1-3 arrive at the destination at the same. Furthermore, if passengers increase, bus 4 runs with buses 1-3. When the number of passengers is less than the capacity, bus 2 stops at the origin and waits for a chance. Running buses return to the origin at the same time.

The increase (decrease) of buses is determined by the number of passengers waiting at the origin. However, it takes a time to arrange an additional bus. Therefore, a delay occurs when buses increase or decrease according to the number of passengers waiting at the origin. The delayed increase (decrease) of buses is taken into account in the bus transportation system. The delay is defined as τ .

The capacity of a single bus is defined as $F_{\max}$. In order to board all passengers waiting at the origin, $M(n - \tau)$ buses are necessary at trip n . The number of buses is given by

$$M(n - \tau) = 1 + \text{int} \left[\frac{B(n - \tau)}{F_{\max}} \right], \quad (26)$$

where $B(n - \tau)$ is the number of passengers at trip $n - \tau$.

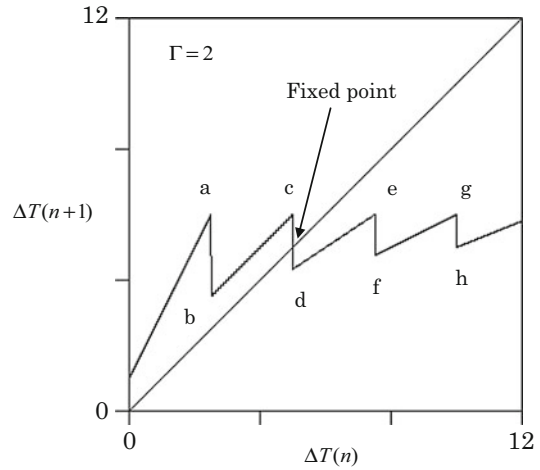
If passengers are distributed uniformly for M buses at the origin, the number of passengers boarding a bus at trip n is given by $B(n)/M(n - \tau)$. Then, the dimensionless arrival time $T(n + 1)$ of buses at the origin at trip $n + 1$ is given by

$$T(n+1) = T(n) + \frac{\Gamma(T(n) - T(n-1))}{M(n-\tau)} + 1. \quad (27)$$

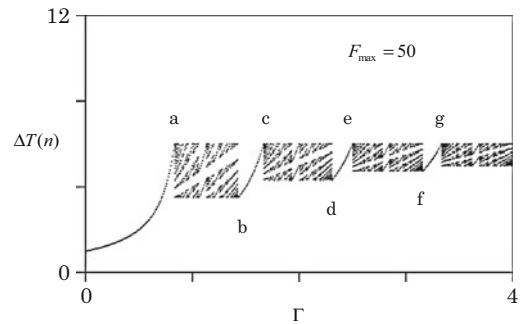
Nonlinear map (27) is the delayed piecewise linear map. Delay τ occurs when buses increase or decrease according to the number of passengers waiting at the origin. When $\tau = 0$, the number of buses at trip n is determined by the number of waiting passengers at trip n . For $\tau = 0$, Eq. 27 is a one-dimensional piecewise linear map. The fixed point is derived analytically. If $\tau = 1$, the increase or decrease of buses is one trip late. Delayed map (27) is a $(1 + \tau)$ -dimensional map. The one-dimensional map changes to $(1 + \tau)$ -dimensional map (27) by introducing the delay of the increase of buses. The characteristics of the high-dimensional map are very complex, and it is difficult to derive the fixed point analytically.

Figure 19 shows the piecewise linear map at $\Gamma = 2$. There is an unstable fixed point in region $c-d$. When the fixed point exists in region $c-d$, two or three buses run in a complex manner because the fixed point is unstable.

Figure 20 shows the plot of dimensionless tour time $\Delta T(n)$ against dimensionless inflow rate Γ for arrival numbers (trip) n from $n = 100$ to $n = 1000$ at $\tau = 0$, and number limit $F_{\max} = 50$. In the region between 0 and point a , a single bus runs with the constant value of tour time, and the tour time agrees with the value of fixed point. In the region between points a and b , a bus or two buses run periodically, and the tour time takes multiple values due to the periodic motion of buses. In the region between points b and c , two buses run with the constant value of tour time, and the tour time agrees with the value of fixed point. In the region between points c and d , two or three buses run periodically, and the tour time takes multiple values due to the periodic motion of buses. In the region between points d and e , three buses run with the constant value of tour time, and the tour time agrees with the value of fixed point. In the region between points e and f , three or four buses run periodically, and the tour time takes multiple values due to the periodic motion of buses. In the region between points f and g , four buses run with the constant value of



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 19 Piecewise linear map at $\Gamma = 2$. There is an unstable fixed point in region $c-d$. When the fixed point exists in the region $c-d$, two or three buses run in a complex manner because the fixed point is unstable



Complex Dynamics of Bus, Tram, and Elevator Delays in Transportation Systems, Fig. 20 Plot of tour time $\Delta T(n)$ against inflow rate Γ at $\tau = 0$ and capacity $F_{\max} = 50$. In the region between 0 and point a , a single bus runs with the constant value of tour time, and the tour time agrees with the value of fixed point. In the region between points a and b , a bus or two buses run periodically, and the tour time takes multiple values due to the periodic motion of buses. In the region between points b and c , two buses run with the constant value of tour time, and the tour time agrees with the value of fixed point. In the region between points c and d , two or three buses run periodically, and the tour time takes multiple values due to the periodic motion of buses

tour time, and the tour time agrees with the value of fixed point. In the region between points g and 4, four or five buses run periodically and the tour time takes multiple values due to the periodic

motion of buses. Thus, the buses behave in a complex manner by increase or decrease of buses.

If there is a delay for arranging additional buses, the dynamic behavior of buses is very complex and shows the regular, periodic, and chaotic motions.

Here, we assumed that all buses start at the same time from the origin and move together. In this case, a bus does not pass over other buses. If passing is allowed, buses run a neck-and-neck race with other buses. The model of increase or decrease buses changes. The passing has an important effect on the bus dynamics by the neck-and-neck race.

Future Directions

Here, we considered the single bus system on the single route and the bus system of a few buses on two routes. In the real transportation system, many buses run simultaneously on multiple routes. In a city, various transportation networks are formed in a complex manner. Men and women go to the destination by using buses, trams, and vehicles. People go from the starting point, through the transfer point, to the destination. One wishes to know how the trip time takes from the starting point, through the transfer point, to the destination. One thinks about where are the transfer points. It is also important to select one proper transport system and route from various number of choices. Here, we presented some transport models of buses, trams, and elevators. The dynamic models of bus transport systems are simple compared to the real transportation system. It is necessary and important to extend the above models to be capable of the real transportation network. In the transportation network, it is important to make a route choice. It is necessary to extend the present models for the public transport model in the network. At the first step to the extension, one will be able to make a dynamic model in which the bus changes are taken into account. By accounting the bus changes at some transfer points, one can extend the nonlinear map models to the coupled map model.

In real traffic, the demand (the arrival rate of the passengers) varies irregularly and largely.

When the demand is stochastic, weak stochasticity does not affect the qualitative features, but strong stochasticity will have an important effect on the stability and dynamics of the system. It will be necessary to investigate the effect of stochastic demand on the bus's motion.

Bibliography

Primary Literature

- Chowdhury D, Desai RC (2000) Steady-states and kinetics of ordering in bus-route models: connection with the Nagel-Schreckenberg model. *Eur Phys J B* 15:375–384
- Dong C, Ma X, Wang B, Sun X (2010) Effects of prediction feedback in multi-route intelligent traffic systems. *Phys A* 389:3274–3281
- Gupta AK, Sharma S, Redhu P (2015) Effect of multi-phase optimal velocity function on jamming transition in a lattice hydrodynamic model with passing. *Nonlinear Dyn* 80:1091–1108
- He DR, Yeh WJ, Kao YH (1985) Studies of return maps, chaos, and phase-locked states in a current-driven Josephson-junction simulator. *Phys Rev B* 31:1359–1373
- Hino Y, Nagatani T (2014) Effect of bottleneck on route choice in two-route traffic system with real-time information. *Phys A* 395:425–433
- Hino Y, Nagatani T (2015) Asymmetric effect of route-length difference and bottleneck on route choice in two-route traffic system. *Phys A* 428:416–425
- Huijberts HJC (2002) Analysis of a continuous car-following model for a bus route: existence, stability and bifurcations of synchronous motions. *Phys A* 308:489–517
- Kerner BS (2004) Three-phase traffic theory and highway capacity. *Phys A* 333:379–440
- Kerner BS (2016a) The maximization of the network throughput ensuring free flow conditions in traffic and transportation networks: breakdown minimization (BM) principle versus Wardrop's equilibria. *Eur Phys J B* 89:199
- Kerner BS (2016b) Failure of classical traffic flow theories: stochastic highway capacity and automatic driving. *Phys A* 450:700–747
- Kerner BS (2017) Breakdown minimization principle versus Wardrop's equilibria for dynamic traffic assignment and control in traffic and transportation networks: a critical mini-review. *Phys A* 466:626–662
- Komada K, Kojima K, Nagatani T (2011) Vehicular motion in 2D city traffic network with signals controlled by phase shift. *Phys A* 390:914–928
- Lämmer S, Gehlsen B, Helbing D (2006) Scaling laws in the spatial structure of urban road networks. *Phys A* 363:89–95
- Li X, Fang K, Peng G (2017) A new lattice model of traffic flow with the consideration of the drivers' aggressive characteristics. *Phys A* 468:315–321

- Masoller C, Rosso OA (2011) Quantifying the complexity of the delayed logistic map. *Philos Trans R Soc A* 369:425–438
- Nagatani T (2000) Kinetic clustering and jamming transitions in a car following model of bus route. *Phys A* 287:302–312
- Nagatani T (2001a) Bunching transition in a time-headway model of bus route. *Phys Rev E* 63:036115-1-7
- Nagatani T (2001b) Interaction between buses and passengers on a bus route. *Phys A* 296:320–330
- Nagatani T (2001c) Delay transition of a recurrent bus on a circular route. *Phys A* 297:260–268
- Nagatani T (2002a) Bunching and delay in bus-route system with a couple of recurrent buses. *Phys A* 305:629–639
- Nagatani T (2002b) Dynamical transition to periodic motions of a recurrent bus induced by nonstops. *Phys A* 312:251–259
- Nagatani T (2002c) Transition to chaotic motion of a cyclic bus induced by nonstops. *Phys A* 316:637–648
- Nagatani T (2002d) Dynamical behavior in the nonlinear-map model of an elevator. *Phys A* 310:67–77
- Nagatani T (2002e) Chaotic and periodic motions of a cyclic bus induced by speedup. *Phys Rev E* 66:046103-1-7
- Nagatani T (2003a) Chaotic motion of shuttle buses in two-dimensional map model. *Chaos Solitons Fractals* 18:731–738
- Nagatani T (2003b) Complex behavior of elevators in peak traffic. *Phys A* 326:556–566
- Nagatani T (2003c) Fluctuation of tour time induced by interactions between cyclic trams. *Phys A* 331:279–290
- Nagatani T (2003d) Transitions to chaos of a shuttle bus induced by continuous speedup. *Phys A* 321:641–652
- Nagatani T (2003e) Complex motions of shuttle buses by speed control. *Phys A* 322:685–697
- Nagatani T (2003f) Dynamical transitions to chaotic and periodic motions of two shuttle buses. *Phys A* 319:568–578
- Nagatani T (2003g) Chaos and headway distribution of shuttle buses that pass each other freely. *Phys A* 323:686–694
- Nagatani T (2003h) Fluctuation of riding passengers induced by chaotic motions of shuttle buses. *Phys Rev E* 68:036107-1-8
- Nagatani T (2003i) Dynamical behavior of N shuttle buses not passing each other: chaotic and periodic motions. *Phys A* 327:570–582
- Nagatani T (2004) Dynamical transitions in peak elevator traffic. *Phys A* 333:441–452
- Nagatani T (2005a) Self-similar behavior of a single vehicle through periodic traffic lights. *Phys A* 347:673–682
- Nagatani T (2005b) Chaos and dynamics of cyclic trucking of size two. *Int J Bifurcat Chaos* 15:4065–4073
- Nagatani T (2006a) Control of vehicular traffic through a sequence of traffic lights positioned with disordered interval. *Phys A* 368:560–566
- Nagatani T (2006b) Chaos control and schedule of shuttle buses. *Phys A* 371:683–691
- Nagatani T (2007a) Clustering and maximal flow in vehicular traffic through a sequence of traffic lights. *Phys A* 377:651–660
- Nagatani T (2007b) Dynamical model for retrieval of tram schedule. *Phys A* 377:661–671
- Nagatani T (2007c) Passenger's fluctuation and chaos on ferryboats. *Phys A* 383:613–623
- Nagatani T (2008) Dynamics and schedule of shuttle bus controlled by traffic signal. *Phys A* 387:5892–5900
- Nagatani T (2011a) Complex motion of shuttle buses in the transportation reducing energy consumption. *Phys A* 390:4494–4501
- Nagatani T (2011b) Complex motion in nonlinear-map model of elevators in energy-saving traffic. *Phys Lett A* 375:2047–2050
- Nagatani T (2012) Delay effect on schedule in shuttle bus transportation controlled by capacity. *Phys A* 391:3266–3276
- Nagatani T (2013a) Dynamics in two-elevator traffic system with real-time information. *Phys Lett A* 377:3296–3299
- Nagatani T (2013b) Nonlinear-map model for control of airplane. *Phys A* 392:6545–6553
- Nagatani T (2013c) Modified circle map model for complex motion induced by a change of shuttle buses. *Phys A* 392:3392–3401
- Nagatani T (2013d) Complex motion of elevators in piecewise map model combined with circle map. *Phys Lett A* 377:2047–2051
- Nagatani T (2013e) Nonlinear-map model for bus schedule in capacity-controlled transportation. *App Math Model* 37:1823–1835
- Nagatani T (2014) Dynamic behavior in two-route bus system with real-time information. *Phys A* 413:352–360
- Nagatani T (2015) Complex motion induced by elevator choice in peak traffic. *Phys A* 436:159–169
- Nagatani T (2016a) Effect of stopover on motion of two competing elevators in peak traffic. *Phys A* 444:613–621
- Nagatani T (2016b) Effect of speedup delay on shuttle bus schedule. *Phys A* 460:121–130
- Nagatani T (2016c) Complex motion of a shuttle bus between two terminals with periodic inflows. *Phys A* 449:254–264
- Nagatani T (2017) Effect of periodic inflow on speed-controlled shuttle bus. *Phys A* 469:224–231
- Nagatani T, Naito Y (2011) Schedule and complex motion of shuttle bus induced by periodic inflow of passengers. *Phys Lett A* 375:3579–3582
- Nagatani T, Tobita K (2012) Effect of periodic inflow on elevator traffic. *Phys A* 391:4397–4405
- Nagatani T, Yoshimura J (2002) Dynamical transition in a coupled-map lattice model of a recurrent bus. *Phys A* 316:625–636
- Naito Y, Nagatani T (2012) Effect of headway and velocity on safety-collision transition induced by lane changing in traffic flow. *Phys A* 391:1626–1635
- O'loan OJ, Evans MR, Cates ME (1998) Jamming transition in a homogeneous one-dimensional system: the bus route model. *Phys Rev E* 58:1404–1421
- Redhu P, Gupta AK (2016) The role of passing in a two-dimensional network. *Nonlinear Dyn* 86:389–399

- Sugiyama Y, Nagatani T (2012) Multiple-vehicle collision induced by a sudden stop. *Phys Lett A* 376:1803–1806
- Sugiyama Y, Nagatani T (2013) Multiple-vehicle collision in traffic flow by a sudden slowdown. *Phys A* 392:1848–1857
- Tobita K, Nagatani T (2012) Effect of signals on two-route traffic system with real-time information. *Phys A* 391:6137–6145
- Tobita K, Nagatani T (2013) Green-wave control of unbalanced two-route traffic system with signals. *Phys A* 392:5422–5430
- Toledo BA, Munoz V, Rogan J, Tenreiro C, Valdivia JA (2004) Modeling traffic through a sequence of traffic lights. *Phys Rev E* 70:016107
- Toledo BA, Cerda E, Rogan J, Munoz V, Tenreiro C, Zarama R, Valdivia JA (2007) Universal and non-universal feature in a model of city traffic. *Phys Rev E* 75:026108
- Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62:1805
- Treiber M, Kesting A, Helbing D (2006) Delays, inaccuracies and anticipation in microscopic traffic models. *Phys A* 360:71–88
- Villalobos J, Toledo BA, Pasten D, Nunoz V, Rogan J, Zarama R, Valdivia JA (2010) Characterization of the nontrivial and chaotic behavior that occurs in a simple city traffic model. *Chaos* 20:013109
- Wahle J, Lucia A, Bazzan C, Klugl F, Schreckenberg M (2000) Decision dynamics in a traffic scenario. *Phys A* 287:669–681
- Wastavino LA, Toledo BA, Rogan J, Zarama R, Muñoz V, Valdivia JA (2008) Modeling traffic on crossroads. *Phys A* 381:411–419

Books and Reviews

- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199–329
- Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
- Kerner BS (2004) *The physics of traffic*. Springer, Heidelberg
- Nagatani T (2002) *The physics of traffic jams*. Rep Prog Phys 63:1331–1386
- Paterson SE, Allan LK (2009) *Road traffic: safety, modeling, and impacts*. Nova Science Publishers, New York

Travel Behavior and Demand Analysis and Prediction

Konstadinos G. Goulias
University of California Santa Barbara, Santa
Barbara, CA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Dynamic Planning Practice
Sustainable and Green Visions
New Research and Technology
The Evolving Modeling Paradigm
Examples of Mathematical Models
Summary
Future Directions
Bibliography

Glossary

Activity-based approach A modeling method that accounts for the interdependent relationships among activities and persons to derive travel demand equations.

Dynamic planning The incorporation of trends, cycles, and feedback mechanisms into a process of actively shaping our future. Desired futures are first defined in terms of performance measures and a combination of forecasting and backcasting methods are used to identify the right paths to follow in achieving these futures.

Microsimulation A method to represent the movement in space and time of the most elementary units of a phenomenon. When applied in traffic engineering the units are vehicles. When applied in travel behavior the units are persons and households. Multi-agent microsimulation allows to also represent

human interaction with each person modeled as an agent.

Travel demand The amount of travel within a time interval such as number of trips in a day, total amount of distance and total amount of travel time, the locations (destinations) visited, the means used to reach these locations, departure time and arrival time of trips, routes followed in reaching these locations, the sequencing and assembly of trips in groups, and the purpose or activity engaged in at the end of each trip.

Definition of the Subject

Transportation modeling and simulation aims at the design of an efficient infrastructure and service to meet our needs for accessibility and mobility. At its heart is good understanding of human behavior that includes the identification of the determinants of behavior and the change in human behavior when circumstances change either due to control (e.g., policy actions), trends (e.g., demographic change), or unexpectedly (e.g., disasters). This is the key ingredient that drives most decisions in transportation planning and traffic operations. Since transportation systems are the backbone connecting the vital parts of a city (a region, a state or an entire country), in-depth understanding of transportation-related human behavior is essential to the planning, design, and operational analysis of all the systems that make a city function.

Understanding human nature requires us to analyze and develop synthetic models of human agency in its most important dimensions and the most elemental constituent parts. This includes, and it is not limited to, understanding of individual evolution along a life cycle path (from birth to entry in the labor force to retirement to death) and the complex interaction between an individual and the anthropogenic environment, natural environment, and the social environment. Travel

behavior research is one aspect of analyzing human nature and aims at understanding how traveler values, norms, attitudes, perception and constraints lead to observed behavior. Traveler values and attitudes refer to motivational, cognitive, situational, and disposition factors determining human behavior. Travel behavior refers primarily to the modeling and analysis of travel demand, based on theories and analytical methods from a variety of scientific fields. These include, but are not limited to, the use of time and its allocation to travel and activities, methods to study this in a variety of time contexts and stages in the life of people, and the arrangement or artifacts and use of space at any level of social organization such as the individual, the household, the community, and other formal or informal groups. This includes the movement of goods and the provision of services having strong interfaces and relationships with the engagement in activities and the movement of persons.

Travel behavior analysis and synthesis can be examined from both objective (observed by an analyst) and subjective (perceived by the human) perspectives in an integrated manner among four dimensions of time, geographic space, social space, and institutional context. In a few occasions the models reviewed here include and integrate time and space as conceived in science with perceptions of time and space by humans in their everyday life. For this reason research includes theory formation, data collection, modeling, inference, and simulation methods to produce decision support systems for policy assessment and evaluation that combine different views of time and space. Another objective of understanding human behavior is conceptual integration. Explanation of facts from different perspectives can be considered jointly to form a comprehensive understanding of people and their groups and their interactions with the natural and built environment. In this way, we may see explanations of human behavior fusing into the same universal principles. These principles eventually will lead to testable hypotheses from different perspectives offering Wilson's, 1998, famous consilience among, for example, psychology, anthropology, economics, the natural sciences, geography, and engineering. Unavoidably this is a

daunting task with many model propositions in the research domain and very few ideas finding fertile ground in applications. The analysis-synthesis path in travel behavior gave us methods that help us understand and predict human (travel) behavior only partially leaving many gaps (Timmermans 2003). However, policy questions are becoming increasingly impossible to address with old tools, a large pool of researchers is actively working on new methods, and many public agencies commenced a variety of tool development projects to fill the travel behavior analysis gaps. To capture these trends, we see modeling examples with ideas from a transdisciplinary viewpoint and contributors to modeling and simulation from a variety of merged backgrounds (e.g., see the evolution of ideas in a sequence of the International Association for Travel Behavior Research conferences – www.public.asu.edu/~rpendyal/iatbr/iatbr_index.htm).

In the next sections the evolving paradigm of modeling and simulation is reviewed in detail and three of its fundamental sources are presented. Through the lens of contemporary planning practice the analytical requirements for modeling and simulation are discussed. Then, these same requirements are refined by examining contemporary visions about the world surrounding us and the theories and technologies we can use to build policy analysis models. This article ends with a section describing the emerging modeling and simulation paradigm, a brief section of mathematical models and closes with a summary.

Introduction

The impressive movement forward of transportation modeling and simulation emerges from three related but distinct sources. The first source is a fundamental change in planning practice that one could name *dynamic planning practice* to indicate the existence of bi-directional time (from the past to the future and from the future to today), as well as, assessment cycles and adjustments taking place within the short term, medium term, and long term horizons. These cycles are also bidirectional in time. This source contains three fundamental directions of practice that are *inventory*

creation and maintenance, strategy measurement and evaluation, forecasting and backcasting. The second source is a vision that generates the substantive problems that we need to solve and the specific policies we need to examine. It is named *sustainable and green visions*. Problems and solutions in this general area motivate and inspire contemporary substance and content of policies throughout the world. One can identify three complementary and mutually strengthening directions in the *economy, environment, and society* that are the three fundamental pillars of sustainability. The third source is the never ending research for improved understanding of the world surrounding us. This source is named *new research and technology* to capture the most important elements of new discovery and new techniques enabling new discovery but also modeling and simulation. Key directions of inquiry within research and technology are *theory building, modeling and simulation, and enabling technologies*.

Dynamic Planning Practice

Dynamic thinking means that time and change are intrinsic in the thought processes underlying planning activities. In the past, assumptions about the existence of a tenable and general equilibrium and our ability to build the infrastructure needed to meet demand did not require careful orchestration of actions. This was radically changed in the industrialized world to meet specific goals using available finite resources to maximize benefits. Together with our inability to build at will and a tendency to the preservation of non-renewable resources (e.g., land and open space, fossil fuels, time) we are much more motivated to think strategically and to consider in a more careful way the performance of the overall anthropogenic system as we plan, design, operate, and manage transportation systems. Any action of this type, however, requires that we have a detailed and accurate picture of our facilities, their interconnectedness, their status within the hierarchy of movements, their conditions, and their evolving role. An accurate and more complete picture like this is called an *inventory* herein.

Many planning activities at all geographical levels are preceded by data gathering steps of identifying all the sources of data and information about the specific study area's transportation system and its relationship with the rest of the world. These inventories include the typical information about the resident population – demographics and employment, land available and land uses, economic development and growth, and so forth. It is worth pointing out the inventory contains data and relationships within the geographic area of interest (region) but also the region's relationship with other areas with which substantial flow of people, goods, and communication takes place. Inventories may also include data and information about cultural and historical factors. For example, state-wide plans identify a variety of corridors as buffers of land and communities around major routes of the movement of people and goods. Some of these routes were created centuries ago when pioneers were still exploring uncharted lands. These routes experienced a major change when waterways were the main links among economic and military centers, and they are still evolving. Today these same routes contain as backbones railways, freeways, rivers, and often they surround major distribution locations such as ports and airports. Their nature is heavily influenced by their historical and cultural context.

Travel behavior analysts are familiar with inventories created for the regional long range plans, which subdivide the study area in traffic analysis zones with data from the Decennial Census suitably reformatted and packaged for use in a specific application (i.e., the long range regional plan). Then, additional data are assigned to these same subdivisions to build a richer context for modeling and simulation. Thus, the inventory for a typical long range plan is an electronic map of where people live and work, the network(s) that connect different locations, availability of different modes on each segment of the network, as well as information about travel network performance (e.g., link capacities, speeds on links, congestion, and connectivity). Today the tool of choice for data storage and visualization is a Geographic Information System (GIS).

One of the thorniest problems within this context is maintaining an up to date inventory (e.g., characteristics of the population in each zone, presence of certain types of businesses, location and characteristics of intermodal facilities). This is a particularly important issue for periods in between decennial censuses. Year to year updates are very often required to provide “fresh” data. Many of these updates are becoming widely available and much less expensive than in the past. For example, the inventory of the highway network, with suitable additions and improvements, is available from the same private providers of invehicle navigation systems. In a similar way, inventories of businesses and residences can also be purchased from vendors. Census data, however, are required even when one uses data from private providers because they contain complementary data (e.g., the age distribution of the resident population) and they tend to provide wider coverage of a country. Although the need for inventories is undoubtedly extremely important many important issues are yet to be resolved. This is the core issue of two Transportation Research Board (TRB) conference proceedings on the National Household Travel Survey <http://www.trb.org/Conferences/NHTS/Program.pdf> and the US Census and the Census American Community Survey <http://www.trb.org/conferences/censusdata/>). Examples of unresolved issues include levels of detail we should use in updating the data we have, treatment of errors in the data and model sensitivity to these errors, frequency of data updates and treatment of missing data, and questions about merging different databases. Obviously, the answers to these questions are in the form of “it depends.” It depends on the budget (time and money) available, consequences of errors in the data, and the use of models in decision making. In fact, one particular type of data collection is strategy measurement where some of these questions become even more important. We turn now to the second dimension in the dynamic planning practice which is about strategy and performance.

Strategic planning and performance-based planning changed the way we plan for the future. This has been a 20 year long process in the United

States as its transportation policy at the Federal, State, and Metropolitan levels is shaped by three consecutive legislative initiatives (ISTEA, TEA-21, and SAFETEA-LU). Under all three legislative frameworks and independently of role, location and perceived need for investment, the overall goal of funding allocation has been to maximize the performance of the transportation system in its entirety and avoid major new infrastructure building initiatives. As a result, planning practice at the Federal, State, and local levels is becoming heavily performance based and designed in a way that motivates the measurement of policy and program outcomes and judging these outcomes for funding allocation. Two examples of performance-based planning are the Program Assessment Rating Tool (PART) at the federal level and performance-based transportation planning at the state level. PART is used to assess the management and performance of individual programs from homeland security to education, employment, and training. This is a tool that offers assessments about programs based on 25 questions divided into sections. For each program a tailored analysis yields summaries that receive a rating from 0 to 100 ranging from ineffective to effective (US Government 2006)). In a different way but in the same spirit many states have created long range plans that are strategic and they measure transportation performance. Yearly evaluative updates are also used for a state’s strategic transportation plan. After a comprehensive public involvement campaign a few themes capturing the desires of the resident population are first identified. To these themes technical requirements based on planners and agency inputs are added, a large number of objectives are created and then a variety of measures of performance are developed. These measures are given target levels that evolve over time to a desired future performance for the entire state and for a finite number of corridors of statewide significance. Yearly evaluations contain measures of target achievement and they should be used to guide an agency in its investments. The interface with regions is also included in this performance-based framework. Many infrastructure improvement projects in the US are selected from lists of

projects that regions (called Metropolitan Planning Organizations) submit to their state to be included in a list of projects in the Transportation Improvement Program (TIP) and become candidates for funding. Under statewide performance-based planning, these projects are evaluated with respect to their contribution in meeting the statewide performance measures and in some states the performance measures of the relevant corridor (National Cooperative Highway Research Program 2000). Although these examples are far ranging in time and space, they contain operations components and yearly evaluations that: a) require data collection, modeling, and simulation at finer spatial and temporal scales than their counterpart planning feedbacks used in the long range transportation planning practice, and b) need a method that is able to coordinate the short, medium, and long term impacts. Emerging from these considerations are questions about the types of consistency we need among geographic scales for planning and operations actions to perform evaluations, policy requirements for coordination among planning activities to ensure consistency, need for suitable methods to coordinate smaller projects in broader contexts (either of policy assessment or geographical area), development of tools required to perform measurement of impacts and program evaluation at the newly defined assessment cycles, and optimal planning activity with evaluation methods. Only a few solutions to the issues above are offered by contemporary projects such as the TRANSLAND project (Grieving and Kemper 1999). Within the context of integration between land use and transportation planning and the context of the European Union some of the conclusions include a call to strengthen regional plans, a stronger emphasis on public transport, strategic planning involving all actors, and the packaging of policies aiming at the same objectives. These themes are very similar to statewide and US Federal and European Union levels of planning. Very little, however, is said about the assessment methods and the choices we make in impact estimation. Performance assessment and evaluation of program effectiveness require the use of the inventory discussed before and a battery of models to

forecast future expectations as well as to identify the actions required today to achieve desired futures.

As illustrated later in this article a new approach emerges in which models of discrete choice are applied to individual decision makers that are then used to (micro)simulate most of the possible combinations of choices in a day. The result is in essence a synthetic generation of traveling for the entire population. When the microsimulation also includes activities and duration at activity locations it becomes a synthetic schedule. In parallel, for forecasting purposes a synthetic population is first created for each land subdivision with all the relevant characteristics and then models are applied to the residents of each subdivision to represent areawide behavior. Changes are then imposed on each individual as a response to policies and predictive scenarios of policy impacts are thus developed. The evolution of individuals, their groups, and the entire study area can be used for trend analysis that includes details at the level of decision makers (either for passenger travel and/or for freight). In addition, progression in time happens from the present to the future and one could identify paths of change by individuals and groups if the application has been designed in the proper way (e.g., keeping detailed accounting of individuals as they move in time, using models that are designed for transitions over time and so forth). In a forecasting setting progression in time follows calendar time, temporal resolution is most often a year, and the treatment of dynamics is an one-way causal stream to the future.

Within the broader study of futures, forecasting is the method we use to develop *projective scenarios*. Performance-based planning, however, requires tools that can extrapolate from future performance targets the actions required today to reach them. In essence we also need *prospective studies* that start from a desirable future and move backwards to identify specific actions that will lead us to that prospect. *Backcasting* was invented in a study of future energy options by (Robinson 1982), to do exactly this through a participatory process. Scenarios in backcasting are the “images” of the future and the possible paths

that will take us to that future. A typical application includes the stages shown in Table 1. An open question, however, remains with respect to scenario construction and assessment. This is particularly important when one considers the serious issues we face with inadequate design of experiments/trials in the forecasting setting. Forecasting and backcasting have some important differences in their objectives. On one hand forecasting is employed to identify likely futures and to develop methods to help us identify small changes in our policies. It is also a method to extrapolate past trends into the future and possibly identify paths of changes that are heavily influenced by habit and inertia. Backcasting, on the other hand is designed to discover new ways to build desirable futures. It is perfectly aligned with strategic planning and it is a better suited method for developing a program of conditions to meet targets. Many of the models developed to date are designed for forecasting applications (either to inform the design of forecasting model systems or to create necessary components in the model systems). Yet, planning practice is moving towards strategy development and therefore needs model components that fit within a backcasting scenario building (see the reversed four-step model in Miller and Demetsky (1999), and its neural network

implementation in Sadek et al. (2002) and the participatory tools in California (<http://www.sacregionblueprint.org/> – accessed May 2007).

Sustainable and Green Visions

Policy actions also view the world surrounding us as an integral ecosystem placing more emphasis on its overall survival by examining direct and indirect effects of individual policy actions and entire policy packages or programs (see the examples in Meyer and Miller 2001). This trend is not limited to transportation. Lomborg (2001), shows that a sustainable and green vision encompasses the entire range of human activity and the entirety of the ecosystem we live in. Although these are good news, because the approach enables analyzes and policies that are consistent in their vision about futures, comprehensive views also reveal that the pace of economic growth and development is in clear conflict with the biological pace of evolution with unknown consequences (Tiezzi 2003) strengthening the view that more comprehensive analytical frameworks are required.

In fact, one of the most recent studies on research needs, which addresses the transportation and environment relationship by the Transportation Research Board of the National

Travel Behavior and Demand Analysis and Prediction, Table 1 Backcasting schema

Content	Method	
Determine objectives, purpose of the analysis, temporal, spatial and substantive scope of the analysis, decide the number and type of scenarios. Identify endogenous and exogenous variables	Problem orientation with technical representatives and stakeholders	
Specify goals, constraints and targets for each scenario analysis and exogenous variables	Stakeholder creativity workshop and brainstorming sessions	
Describe present system (building and updating of inventories), patterns and trends. Define processes, their actors, and determinants of outcomes. Identify exogenous variables and inputs to scenario analysis.	Scenario development by technical experts	
Scenario analysis. Select suitable approach, analyze system evolution at end time points and intermediate time points, develop scenarios, iterate to make sure all components are consistent/coherent	Scenario assessment by technical experts and stakeholders	
Undertake impact analysis. Consolidate scenario results. Analyze social, economic and environmental impacts. Compare results of the last with targets, iterate analysis with any other step as required to ensure consistency between goals and results	Backcasting workshops and stakeholder consultation (repeat to follow the iterations)	
Implement Policy Actions		

Academies (Transportation Research Board 1999, 2002), expands the envelope to incorporate ecology and natural systems and addresses human health in a more comprehensive way than in the past reiterating the urgency to address unresolved issues about environmental damage. As a result, we also experience a clear shift to policy analysis approaches that have an expanded scope and domain and they are characterized by explicit recognition of transportation system complexity and uncertainty.

Reflecting all this, *sustainable transportation* is now often used to indicate a shift in the mentality of the community of transportation analysts to represent a vision of a transportation system that attempts to provide services that minimize harm to the environment. In fact, in one of the most comprehensive reviews of policies in North America, Meyer and Miller (2001), contrast the non-sustainable to the sustainable approaches. They provide a compelling argument about the change in these policies and pathways toward a more sustainable path. In the US during the past 20 years, the need, to examine these new and more complex policy initiatives, has also become increasingly pressing due to the passage of a series of legislative initiatives (Acts) and associated Federal and State regulations on transportation policy, planning, and programming. The multi-modal character of the new legislation, its congestion management systems and the taxing air quality requirements for selected US regions have motivated many new forecasting applications that in the early years were predominantly based on the Urban Transportation Planning System and related processes but during the last 5 years motivated a shift to richer conceptual frameworks. In point of fact, air quality mandates motivated impact assessments of the so called transportation control measures and the creation of statewide mobile source air pollution inventories (Goulias et al. 1993; Loudon and Dagang 1994; Stopher 1994) that require different analytical forecasting tools than in any pre-1990 legislative initiatives (Niemeier 2003). An added motivation is also lack of substantial funding for transportation improvement projects and a shift to charge the firms that benefit the most from transportation

system improvements creating a need for impact fee-assessment for individual private developers. These assessments create the need for higher resolution in the three dimensions of geography (space), time (time of day), and social space (groups of people with common interests and missions, households, individuals) used in typical regional forecasting models but also the domain of jurisdictions where major decisions are made. They also create a pressing need for interfaces with traffic engineering simulation tools that are approved and/or endorsed in legislation (for examples see Paaswell et al. (1992)). Another push for new tools is the assessment of technologies under the general name of Intelligent Transportation Systems (i.e., bundles of technological solutions in the form of user services attempting to solve chronic problems such as congestion, safety, and air pollution). Natural and anthropogenic tragic recent events are adding requirements for modeling and simulation and urgency in their development and implementation as well as more detail in time and space (Henson and Goulias 2006).

As Garrett and Wachs (1996), discuss in the context of a lawsuit against a regional planning agency in the Bay Area, traditional four-step regional simulation models (Creighton 1970; Hutchinson 1974; Ortuzar and Willumsen 2001) are outpaced by the same legislative stream of the past 20 years that defined many of the policies described above. Unlike the “energy crisis” of the 1970s, the urgency and timeliness of modeling and simulation is becoming more urgent, more complex, and requires an “integrated” approach. Under these initiatives, forecasting models, in addition to long-term land use trends and air quality impacts, need to also address issues related to technology use and information provision to travelers in the short and medium terms. Similarly, the European Union focuses on issues such as: increasing citizen participation, intra-European integration, decentralization, deregulation, privatization, environmental concerns, mobility costs, congestion management by population segments, and private infrastructure finance (see van der Hoorn 1997). Table 2 provide an overview of policy tools that are loosely ordered from the

Travel Behavior and Demand Analysis and Prediction, Table 2 Examples of policy tools

Type of policy tool	Brief description	Source of information ^a
Land use growth and management programs	Legislation that controls for the growth of cities in sustainable paths	www.smartgrowth.org , www.awcnet.org www.fhwa.dot.gov/planning/ppasg.htm www.compassblueprint.org
Land use design and attention to neighborhood design for non-motorized travel	Similar to the previous but with attention paid to individual neighborhoods	www.sustainable.doe.gov/landuse/luothtoc.shtml www.planning.dot.gov/Documents/DomesticScan/domscan2.htm
City annexations and spheres of influence	City boundaries are divided into incorporated, within the sphere of influence, and external to manage growth	http://countypolicy.co.la.ca.us/BOSPolicyFrame.htm www.ite.org/activeliving/files/Jeff_Summary.pdf
Accelerated retirement of vehicles programs	Programs to eliminate high emitting and older technology vehicles	ntl.bts.gov/DOCS/SCRAP.html
Public involvement and education programs	Programs aiming at defining goals based on the public's desires	www.fhwa.dot.gov/reports/pitttd/contents.htm
Health promoting programs	Programs that promote physical activity in travel to benefit health	www.activelivingbydesign.org
Safety measures	A process to incorporate safety considerations in transportation planning	tmip.fhwa.dot.gov/clearinghouse/docs/safety/ www.fhwa.dot.gov/planning/scp/ www.safetyanalyst.org/
Emission control, vehicle miles traveled, and other fee programs (including carbon taxes and trading)	Programs that shift taxation from traditional sources towards pollutant emissions and natural- resource depletion agents	www.fresh-energy.org/ www.fhwa.dot.gov/environment/ www.fightglobalwarming.com/
Congestion pricing and toll collection programs	A premium is charged to travelers that wish to travel during the most congested periods	www.vtpi.org/london.pdf
Parking fee management	Parking pricing used as a tool to restrict access by space and time	www.gmu.edu/depts/spp/programs/parkingTaxes.pdf
Non-motorized systems	Programs to support walking and biking	www.vtpi.org/tdm/tdm25.htm www.psrc.org/projects/nonmotorized
Telecommuting and Teleshopping	The employment of telecommunications to substitute-complement-enhance travel	www.telework-mirti.org www.vtpi.org/tdm/tdm43.htm
Flexible and staggered work programs	Programs that change the workweek of individuals and firms	www.its.dot.gov/JPODOCS/REPTS_PR/13669/section05.htm
Goods movements (freight) programs to improve operations	A variety of programs to facilitate and minimize the damage for freight movement	ntl.bts.gov/DOCS/harvey.html
Highway system improvements in traffic operations and flow	Improved data collection, monitoring, and traffic management	www.transportation.ite.org/mega/default.asp

(continued)

Travel Behavior and Demand Analysis and Prediction, Table 2 (continued)

Type of policy tool	Brief description	Source of information ^a
Intelligent Transportation Systems (ITS)	Use of telecommunications and information technology to manage and control travel	www.itsa.org/ www.ertico.com/ www.its.dot.gov/index.htm/
Special event planning and associated traffic management	Enhanced procedures to handle the demands of a special event	tmcpsf.ops.fhwa.dot.gov/cfprojects/new_detail.cfm?id=32xxxnew=0
Security preparedness through metropolitan planning processes	A process to incorporate safety considerations in transportation planning	www.planning.dot.gov/Documents/Securitypaper.htm
Individualized marketing techniques with improved information and communication with the "customer"	Public programs to provide personal help in changing travel behavior in favor of environmentally friendly modes	www.local-transport.dft.gov.uk/travelplans/index.htm http://www.travelsmart.gov.au/

^aAccessed May 2007

longer term of land use and governance to medium and shorter term operational improvements depending on the lag time required for their impacts to be realized.

These policy initiatives place more complex issues in the domain of regional policy analysis and forecasting and amplify the need for methods that produce forecasts at the individual traveler and her/his household levels instead of the traffic analysis zone level. In addition to the long range planning activities and the typical traffic operations management activities, analysts and researchers in planning need to also evaluate the following: a) traveler and transportation system manager information provision and use (e.g., location based services, smart environments providing real time information to travelers, vehicles, and operators); b) combinations of transportation management actions and their impacts (e.g., parking fee structures and city center restrictions, congestion pricing), and c) assessment of combinations of environmental policy actions (e.g., carbon taxes and information campaigns about health effects of ozone).

To perform all this we need tool that also have forecasting and backcasting capabilities that are more accurate and detailed in space and time. In fact, planning initiatives are moving toward parcel by parcel analysis and yearly assessments. It is also conceivable that we need separate analyzes

for different seasons of a year and days of the week to capture seasonal and within a week variations of travel. Echoing all this and in the context of the Dutch reality Borghers et al. (1997) have identified five information need domains that the new envisioned policy analysis models will need to address and they are (in a modified format from the original list):

1. social and demographic trends that may produce a structural shift in the relationship between places and time allocation by individuals invalidating existing travel behavior model systems;
2. increasing scheduling and location flexibility and degrees of freedom for individuals in conducting their every day business leading to the need to consider additional choices (e.g., departure time from home, work at home, shopping by the internet, shifting activities to the weekend) in modeling travel behavior;
3. changing quality and price of transport modes based on market dynamics and not on external to the travel behavior policies (e.g., the effect of deregulation in public transport);
4. shifting of attitudes and potential cycles in the population outlook about travel options; and
5. changing scales/jurisdictions (scale is the original term used to signify the different

jurisdictions) – different policy actions in different sectors have direct and indirect effects on transportation and different policy actions in transportation have direct and indirect effects in the other sectors (typical example in the US is the welfare to work program).

The first substantive implication of all these considerations is an expanded envelope of modeling and simulation. Many processes that were left outside the realm of transportation modeling and simulation need to be included as stages of the travel model system. One notable example is the inclusion of *residential location choice*, *work location choice*, and *school location choice* to capture the spatial distribution and relative location of important anchor points on travel behavior and to also capture the impact of transportation system availability and level of service on these choices. In this way when implemented policies lead to improved level of service and the relative attractiveness of locations change, shifts in residential location, work location, and possibly school location can be incorporated as impacts of transportation. A similar treatment is needed for *car ownership and car type choices* of households or *fleet sizes and composition* for firms. These correlated choices are expressed as functions of parking availability, energy and other costs and level of service offered by the transportation system (highway and transit). To account for other resources and facilities available for household travel we also need to consider processes for *driver's licensing*, acquiring of *public transportation subscription (passes)*, and participation in *car sharing programs*. In this way, variables of car availability and public transportation availability in households can be used as determinants of travel behavior. Similar treatment is required for policies that change attitudes, perceptions and knowledge about travel options.

To address some of the policies of Table 2, we need to transition to a domain that contains a variety of outputs that include shares of program participation, sensitivity to accessibility and prices, and the usual indicators of travel on networks using input variables from the processes and behaviors discussed up to this point.

Although the number of vehicles per hour per lane is the typical input of traffic operations software, a variety of other variables such as speeds on network links and types of vehicles are also needed for other models such as emissions estimation.

Ideally longer term social, economic, demographic, resource/facilities, and circumstances of people should be converted into yearly schedules identifying periods of vacation, workdays, special occasions, and so forth. These in turn should lead to weekly schedules separating days during which people stay at home from days during which people go to work and days during which they run errands and/or engage in other non-work and non-school related activities. In this way patterns of working days versus not working days can be derived in a natural (con)sequence. As we will see in a later section, a fundamental leap of faith intervenes in practice and converts all this background information into a representative day that is used to create a more or less complete sequence of activities and trips with their destinations and modes used.

In this way decisions and choices people make are organized along the time scale in terms of the time it takes for these events to occur and their implications. For example, decisions about education, careers and occupation, and residential and job location are considered first and they condition everything that happens next. These should be formulated in terms of one or more life course long projects and not represented by a cross-sectional choice model. Similarly, decisions about yearly school and work schedules that determine work days and vacation days in a year should also be modeled as a stream of interrelated choices. Conditional on all this are the daily schedules of individuals and the myriad of decisions determining a daily schedule, which are modeled in much more detail and paying closer attention to the mutual dependency among the different facets of a within a day schedule. The next section explores this further in the context of research and enabling technology. A section on mathematical models later in this article shows the beginning of a new way in modeling a simulation that emphasizes human interaction.

New Research and Technology

The planning and policy analysis discussion identified many requirements for modeling and simulation. Planning and policy expanded the context of travel behavior models to entire life paths of individuals and for this reason a more general modeling framework is emerging. In fact, modeling made tremendous progress toward a comprehensive approach to, in essence, build simulated worlds on computer enabling the study of complex policy scenarios. Although, passenger travel received the bulk of the attention, similar contributions to new research and technology are found in modeling the movement of goods (Southworth 2003; Stefan et al. 2005). The emerging framework, although incomplete, is rich in the directions taken and potential for scientific discovery, policy analysis, and more comprehensive approaches in dealing with sustainability issues.

There are four dimensions that one can identify in building taxonomies of simulation models. The first is the *geographic space* and its conditional continuity, the second is the *temporal scale* and calendar continuity, the third is interconnectedness of *jurisdictions*, and the fourth and most important is the set of relationships in *social space* for individuals and their communities. The first dimension, *geographic space* here is intended as the physical space in which human action occurs. This dimension has played important roles in transportation planning and modeling because the first preoccupation of the transportation system designers has been to move persons from one location to another (i.e., overcoming spatial separation). Initial applications considered the territory divided into large areas (traffic analysis zones), represented by a virtual center (centroid), and connected by facilities (higher level highways). The centroids were connected to the higher level facilities using a virtual connector summarizing the characteristics of all the local roads within the zone. As computational power increased and the types of policies/strategies required increased resolution, the zone became smaller and smaller. Today, is not unreasonable to expect software to handle zones that are as small as a parcel of land and transportation

facilities that are as low in the hierarchy as a local road (the centroid becomes the building on a parcel and the centroid connector is the driveway of the unit and they are no longer virtual).

In modeling and simulation we are interested in understanding human action. For this reason in some applications geographic space needs to consider more than just physical features (Golledge and Stimpson 1997, p. 387) moving us into the notion of place and social space (see also below). The second dimension is *time* that is intended here as continuity of time, irreversibility of the temporal path, and the associated artificiality of the time period considered in many models. For example, models used in long range planning applications use typical days (e.g., a summer day for air pollution). In many regional long-range models the unspoken assumption is that we target a typical work weekday in developing models to assess policies. Households and their members, however, may not always (if at all) obey this strict definition of a typical weekday to schedule their activities and they may follow very different decision making horizons in allocating time to activities within a day, spreading activities among many days including weekends, substituting out of home with in home activities in some days but doing exactly the opposite on others, and using telecommunications only selectively (e.g., on Fridays and Mondays more often than on other days). Obviously, taking into account these scheduling activities is by far more complex than what is allowed in existing transportation planning models. The third dimension is *jurisdictions* and their interconnectedness. The actions of each person are “regulated” by jurisdictions with different and overlapping domains such as federal agencies, state agencies, regional authorities, municipal governments, neighborhood associations, trade associations and societies, religious groups, and formal and informal networks of families and friends. In fact, the federal government defines many rules and regulations on environmental protection. These may end up being enforced by a local jurisdiction (e.g., a regional office of an agency within a city). On the one hand, we have an organized way of governance that clearly defines jurisdictions and policy domains (e.g.,

tax collection in the US). On the other hand, however, the relationships among jurisdictions and decision making about allocation of resources does not follow always this orderly governance principle of hierarchy. A somewhat different and more “bottom up” relationship is found in the social network and for this reason requires a different dimension that is the fourth and final dimension named *social space* and the relationships among persons within this space. For example, individuals from the same household living in a neighborhood may change their daily time allocation patterns and location visits to accommodate and/or take advantage of changes in the neighborhood such as elimination of traffic and the creation of pedestrian zones. Depending on the effects of these changes on the pedestrian network we may also see a shift in the within the neighborhood social behavior. In contrast, increase in traffic to surrounding places may create an outcry by other surrounding neighborhoods, thus, complicating the relationships among the residents.

One important domain and entity within this social space is the household. This has been a very popular unit of analysis in transportation planning recognizing that strong relationships within a household can be used to capture behavioral variation (e.g., the simplest method is to use a household’s characteristics as explanatory variables in a regression model of travel behavior). In this way any changes in the household’s characteristics (e.g., change in the composition due to birth, death, or children leaving the nest or adults moving into the household) can be used to predict changes in travel behavior. New model systems are created to study this interaction within a household looking at the patterns of using time in a day and the changes across days and years. It is therefore very important in modeling and simulation to incorporate in the models used for policy analysis interactions among these four fundamental dimensions, which bring us to the next major issue that of scale.

The typical long range planning analysis is usually defined for larger geographical areas (region, states, and countries) and addresses issues with horizons from 10 to 50 years. In

many instances we may find that large geographic scale means also longer time frames applied to wider mosaics of social entities and including more diverse jurisdictions. On the other side of the spectrum issues that are relevant to smaller geographic scales are most likely to be accompanied by shorter term time frames applied to a few social entities that are relatively homogeneous and subject to the rule of very few jurisdictions. This is one important organizing principle but also an indicator of the complex relationships we attempt to recreate in our computerized models for decision support. In developing the blueprints of these models one can choose from a variety of theories (e.g., neoclassical microeconomics) and conceptual representations of the real world that help us develop these models. At the heart of our understanding of how the world (as an organization, a household, a formal or informal group, or an individual human being) works are models of decision making and conceptual representations of relationships among entities making up this world.

Transportation planning applications are about judgment and decision making of individuals and their organizations. There are different settings of decision making that we want to understand. Three of these settings are the travelers and their social units from which motivations for and constraints to their behavior emerge; the transportation managers and their organizations that serve the travelers and their social units, and the decision makers surrounding goods movement and service provision that contain a few additional actors, Southworth (2003). These include land use markets (see www.urbansim.org). Travelers received considerable attention in transportation planning and the majority of the models in practice aim at capturing their decision making process. The remaining settings received much less attention and they are poorly understood and modeled.

Conceptual models of this process are transformed into computerized models of a city, a region, or even a state in which we utilize components that are in turn models of human judgment and decision making (e.g., travelers moving around the transportation network and visiting

locations where they can participate in activities). Models of this behavior are simplified versions of strategies used by travelers when they select among options that are directly related to their desired activities. In some of these models we also make assumptions about hierarchies of motivations, actions, and consequences. Some of these assumptions are explicit (e.g., when deriving the functional forms of models as in the typical disaggregate choice models or the rules in a production system) and in other models these assumptions are implicit.

When designing transportation planning model interfaces for transportation planners and managers we also implicitly make assumptions about the managers' ability to understand the input, agent representation, internal functioning, and output of these computerized models. Our objective is therefore not only to understand travel behavior and build models that describe and predict human behavior but also to devise tools that allow transportation managers to understand the assumed behavior in the models, study scenarios of policy actions, and define and explain policy implications to others. This, in essence, implies that we, the model system designers, create a platform for a relationship between planners and travelers. A similar but more direct relationship also exists between travelers and transportation managers when we design the observation methods that provide the data for modeling but also the data used to measure attitudes and opinions such as travel surveys. In fact, this relationship is studied in much more detail in the survey design context and linked directly to the image of the agency conducting the survey and the positive or negative impression of the travelers about the sponsoring agency (Dillman 2000). Most transportation research for modeling and simulation, however, has emphasized traveler behavior when building surveys and their models neglecting the interface with the planners. The summary of theories below, however, applies to individuals traveling in a network but also to organizations and planners in the sense used by H.A. Simon in his *Administrative Behavior* (Simon 1997).

Rational decision making is a label associated with human behavior that follows a strategy in

identifying the best course of action. In summary, a decision maker solves an optimization problem and identifies the best existing solution to this problem. Within this more general strategy when an operational model is needed and this operational model provides quantitative predictions about human behavior some kind of mathematical apparatus is needed to produce the predictions. One such machinery is the subjective expected utility (Savage 1954) formulation of human behavior. In developing alternative models to SEU Simon (1983) defines four theoretical components:

- A person's decision is based on a utility function assigning a numerical value to each option – *existence and consideration of a cardinal utility function*;
- The person defines an exhaustive set of alternative strategies among which just one will be selected – *ability to enumerate all strategies and their consequences*;
- The person can build a probability distribution of all possible events and outcome for each alternate option – *infinite computational ability*; and
- The person selects the alternative that has the maximum utility – *maximizing utility behavior*.

This behavioral paradigm served as the basis for a rich production of models in transportation that include the mode of travel, destinations to visit as well as the household residence (see the examples in the seminal textbook by Ben-Akiva and Lerman (1985). It served also as the theoretical framework for consumer choice models and for attempts to develop models for hypothetical situations (see the comprehensive book by Louviere et al. (2000). It has also replaced the aggregate modeling approaches to travel demand analysis as the orthodoxy against which many old and new theories and applications are compared and compete with. SEU can be considered to be a model from within a somewhat larger family of models under the label of weighted additive rule (WADD) models (Payne et al. 1993). Real humans, however, may never behave according to SEU or related maximizing and infinitely

computational capability models (Simon labels this the Olympian model, (Simon 1983)). Based on exactly this argument different researchers in psychology have proposed a variety of decision making strategies (or heuristics). For example, Simon created alternate model paradigms under the label of *bounded rationality – the limited extent to which rational calculation can direct human behavior* (Simon 1983, 1997) to depict a sequence of a person's actions when searching for a suitable alternative. The modeled human is allowed to make mistakes in this search giving a more realistic description of observed behavior (see also Rubinstein 1998). Tversky is credited with another stream of decision making models starting with the *lexicographic approach* [86, in which a person first identifies the most important attribute, compares all alternatives on the value of this attribute, and chooses the alternative with the best value on this most important attribute. Ties are resolved in a hierarchical system of attributes. Another Tversky model (Tversky 1969) assumes a person selects an attribute in a probabilistic way and influenced by the importance of the attribute, all alternatives that do not meet a minimum criterion value (cutoff point) are eliminated. The process proceeds with all other attributes until just one alternative is left and that one is the chosen. This has been named the *elimination by aspects strategies* (EBA) model. Later, Kahneman and Tversky (Tversky 1972) developed *prospect theory* and its subsequent version of *cumulative prospect theory* in Tversky and Kahneman (1992) in which a simplification step is first undertaken by the decision maker editing the alternatives. Then, a value is assigned to each outcome and a decision is made based on the sum of values multiplying each by a decision weight. Losses and gains are treated differently. All these alternatives to SEU paradigms did not go unnoticed in transportation research with early significant applications appearing in the late 1980s. In fact, a conference was organized attracting a few of the most notable research contributors to summarize the state of the art in behavior paradigms and documented in Garling et al. (1998). One of the earlier examples using another of Simon's inventions, the *satisficing behavior – acceptance of viable*

choices that may not be optimal – is a series of transportation-specific applications described in Mahmassani and Herman (1990). Subsequent contributions continue along the path of more realistic models and the most recent example, discussing a few models, by Avineri and Prashker (2003), uses cumulative prospect theory giving a preview of a movement toward more realistic travel behavior models. As Garling et al. (1998) and Avineri and Prashker (2003) point out, these paradigms are not ready for practical applications, contrary to the Mahmassani and colleagues efforts that have been applied, and additional work is required to use them in a simulation framework for applications. Another aspect is *contextual adaptation*. Payne et al. (1993) provide an excellent review of decision making models and their differentiating aspects. They also provide evidence that decision makers *adapt* by switching between decision making paradigms *to the task and the context of their choices*. They also make mistakes and they may also fail to switch strategies. As Vause (1997) discusses to some length transportation applications are possible using multiple decision making heuristics within the same general framework and employing a production system approach (Newell and Simon 1972). A key consideration, however, that has received little attention in transportation is the definition of context within which decision making takes place. Recent production systems (Arentze and Timmermans 2000) are significant improvements over past simulation techniques. However, travelers are still assumed to be passive in shaping the environment within which they decide to act (action space). This action space is viewed as largely made by constraints and not by their active shaping of their context. Goulias (2001, 2003) reviews another framework from human development that is designed to treat decision makers in their active and passive roles and explicitly accounts for mutual influence between an agent (active autonomous decision maker) and her environment.

Transportation modeling and simulation experienced a few tremendously innovative and progressive steps forward. Interestingly these key innovations are from nonengineering fields but

very often transferred and applied to transportation systems analysis and simulation by engineers. These are listed here in a somewhat sequential chronological order merging technological innovations and theoretical innovations. At exactly the time that the Bay Area Rapid Transit system was studied and evaluated in the 1960s, Dan McFadden (the Year 2000 Nobel Laureate in Economics) and a team of researchers produced practical mode choice regression models at the level of an individual decision maker (see <http://emlab.berkeley.edu/users/mcfadden/> – accessed June 2007). The models are based on random utility maximization (of the SEU family) and their work opened up the possibility to predict mode choice rates more accurately than ever before. These models were initially named *behavioral travel-demand models* (Stopher and Meyburg 1976) and later the more appropriate term of *discrete choice models* (Ben-Akiva and Lerman 1985) prevailed. Although restrictive in their assumptions, these models are still under continuous improvement and they have become the standard tool in evaluating discrete choices. Some of the most notable and recent developments advancing the state of the art and practice are:

- Better understanding of the theoretical and particularly behavioral limitations of these models (Gärling et al. 1998; Golledge and Gärling 2003; McFadden 1998);
- more flexible functional forms that resolve some of the problems raised in Williams and Ortuzar (Williams and Ortuzar 1982) allowing for different choices to be correlated when using the most popular discrete choice regression models (Bhat 2000, 2003; Koppelman and Sethi 2000);
- combination of revealed preference, stated choices by travelers, with stated preferences and intentions, answers to hypothetical questions by travelers, availability of data in the same choice framework to extract in a more informative way travelers willingness to use a mode and willingness to pay for a mode option (Ben-Akiva and Morikawa 1989; Louviere et al. 2000). This latter “improvement” enables

us to assess situations that are impossible to build in the real world;

- computer-based interviewing and laboratory experimentation to study more complex choice situations and the transfer of the findings to the real world (Mahmassani and Jou 1998). This direction, however, is also accompanied by a wide variety of research studies aiming at more realistic behavioral models that go beyond mode choice and travel behavior (Golledge and Gärling 2003); and
- expansion of the discrete choice framework using ideas from *latent class models* with covariates that were first developed by Lazarsfeld in the 1950s and their estimation finalized by Goodman in the 1970s (see the review in Goodman (2002), and discrete choice applications in Bockenholt (2002)). This family of models was used in Goulias (1999) to study the dynamics of activity and travel behavior and in the study of choice in travel behavior (Ben-Akiva et al. 2002).

As mentioned earlier the rational economic assumption of the maximum utility model framework (that underlies many but not all of the disaggregate models) is very restrictive and does not appear to be a descriptive behavioral model except for a few special circumstances when the framing of decisions is carefully designed (something we cannot expect to happen every time a person travels on the network). Its replacement, however, requires conceptual models that can provide the types of outputs needed in regional planning applications. A few additional research paths, labeled as *studies of constraints*, are also functioning as gateways into alternate approaches to replace or complement the more restrictive utility-based models. A few of these models also consider knowledge and information provision to travelers. The first aspect we consider is about the choice set in discrete choice models. Choice set is the set of alternatives from which the decision maker selects one. These alternatives need to be mutually exclusive, exhaustive, and finite in number (Train 2003). Identification, counting, and issues related to the alternatives considered have motivated considerable research

in choice set formation (Horowitz 1991; Horowitz and Louviere 1995; Richardson 1982; Swait and Ben-Akiva 1987a, b). Key threat to misspecification of the choice set is the potential for incorrect predictions (Thill 1992). When this is an issue of considerable threat as in destination choice models where the alternatives are numerous, a model of choice set formation appears to be the additional burden (Haab and Hicks 1997). Other methods, however, also exist and they may provide additional information about the decision making processes. Models of the processes can be designed to match the study of specific policies in specific contexts. One such example and a more comprehensive approach defining the choice sets is the situational approach (Brög and Erl 1980). The method uses in depth information from survey respondents to derive sets of reasons for which alternatives are not considered for specific choice settings (individual trips). This allows separation of analyst observed system availability from user perceived system availability (e.g., due to misinformation and willingness to consider information). This brings us to the duality between “objective choice attributes” and “subjective choice attributes.” Most transportation applications, independently of the decision making paradigm adopted, assume the analysts (modelers) and the travelers (modeled) measured attributes to be the same. Modeling the process of perceived constraints may be far more complex when one considers the influence of the context within which decisions are made. Golledge and Stimpson (pp. 33–34 in (Golledge and Stimpson 1997)) describe this within a conceptual model of decision making that has a cognitive feel to it. They also link the situational approach to the activity-based framework of travel extending the framework further (pp. 315–328 in (Golledge and Stimpson 1997)).

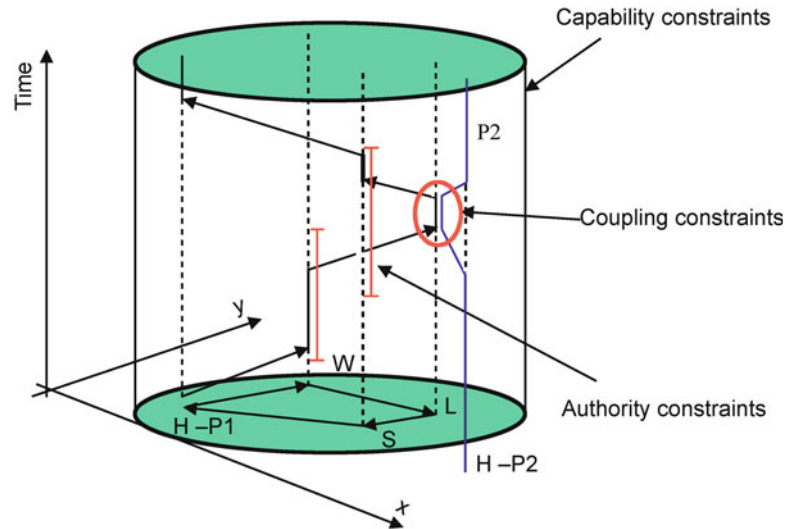
Chapin’s research (Chapin 1974), providing one of the first comprehensive studies about time allocated to activity in space and time, is also credited for motivating the foundations of activity-based approaches to travel demand analysis. His focus has been on the propensity of individuals to participate in activities and travel linking their patterns to urban planning. In about

the same period Becker also developed his theory of time allocation from a household production viewpoint (Becker 1976) applying economic theory in a non-market sector and demonstrating the possibility of formulating time allocation models using economics reasoning (i.e., activity choice). In parallel another approach was developing in geography and Hagerstrand’s seminal publication on time geography (Hagerstrand 1970) presents the foundations of the approach. The idea of constraints in the movement of persons was taken a step further by this time-geography school in Lund. In that framework, the movement of persons among locations can be viewed as their movement in space and time under external constraints. Movement in time is viewed as the one way (irreversible) movement in the path while space is viewed as a three dimensional domain. It provides the third base about *constraints* in human paths in time and space for a variety of planning horizons. These are *capability constraints* (e.g., physical limitations such as speed); *coupling constraints* (e.g., requirements to be with other persons at the same time and place); and *authority constraints* (e.g., restrictions due to institutional and regulatory contexts such as the opening and closing hours of stores). Figure 1 provides a pictorial representation in space and time of a typical activity travel pattern of two persons (P1 and P2) and the three types of constraints. H indicates home, W indicates work, L indicates leisure, and S indicates shopping.

Cullen and Godson (1975) also reviewed by Arentze and Timmermans (2000) and Golledge and Stimpson (1997) appear to be the first researchers attempting to bridge the gap between the motivational (Chapin) approach to activity participation and the constraints (Hagerstrand) approach by creating a model that depicts a routine and deliberated approach to activity analysis. The Cullen and Dobson study also defined many terms often used today in activity-based approaches. For example, each activity (stay-home, work, leisure, and shopping) is an episode characterized by start time, duration, and end time. Activities are also classified into fixed and flexible and they can be engaged alone or with others. Moreover, they also analyzed sequencing

Travel Behavior and Demand Analysis and Prediction, Fig. 1

A two-person (P1 and P2) activity-travel pattern and the time and space limits imposed by constraints (Pribyl 2004)



of activities as well as pre-planned, routine, and on the spur of the moment activities. Within this overall theoretical framework is the idea of a project which according to Golledge and Stimpson (1997), is a set of linked tasks that are undertaken somewhere at some time within a constraining environment (pp. 268–269). This idea of the project underlies one of the most exciting developments in activitybased approaches to travel demand analysis and forecasting because seemingly unrelated activity and trip episodes can be viewed as parts of a “big-picture” and given meaning and purpose completing in this way models of human agency and explaining resistance to change behavior.

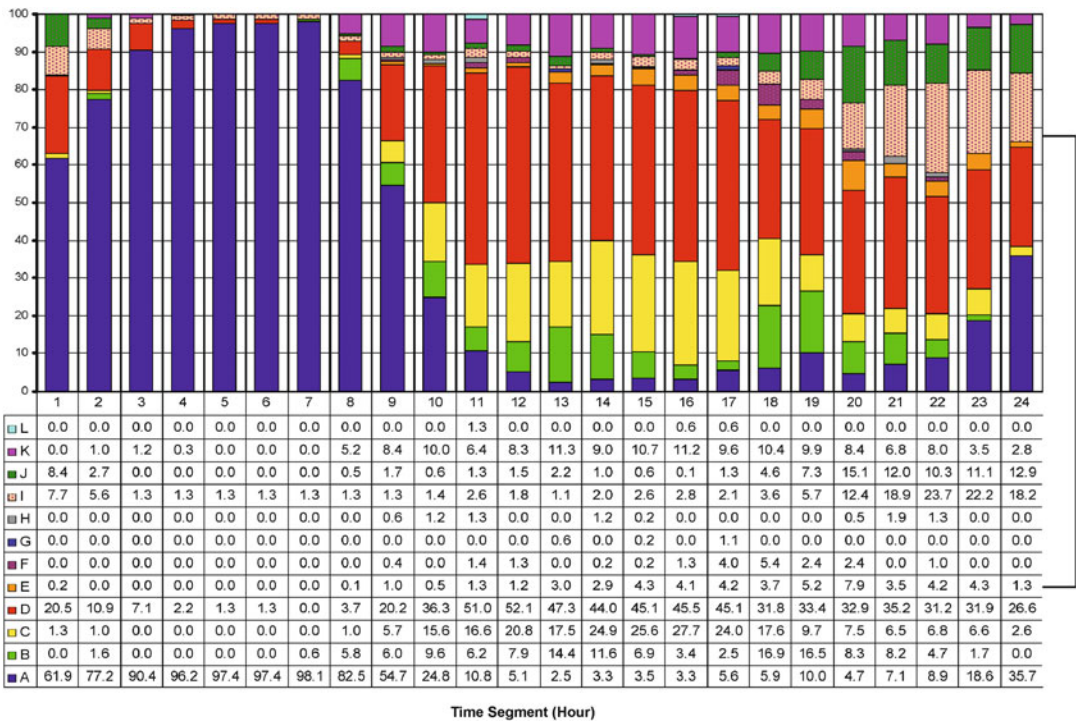
Most subsequent contributions to the activity-based approach emerge in one way or another from these initial frameworks with important operational improvements (for reviews see Arentze and Timmermans 2000; Bhat and Koppelman 1999; Kitamura 1988; McNally 2000). The basic ingredients of an activity based approach for travel demand analysis (Arentze and Timmermans 2000; Jones et al. 1990) are:

- (a) explicit treatment of travel as derived demand (Manheim 1979), i.e., participation in activities such as work, shop, and leisure motivate travel but travel could also be an activity as well (e.g., taking a drive). These activities are

viewed as episodes (i.e., they are characterized by starting time, duration, and ending time) and they are arranged in a sequence forming a pattern of behavior that can be distinguished from other patterns (i.e., a sequence of activities in a chain of episodes). In addition, these events are not independent and their interdependency is accounted for in the theoretical framework;

- (b) the household is considered to be the fundamental social unit (i.e., decision making unit) and the interactions among household members are explicitly modeled to capture task allocation and roles within the household, relationships at one time point and change in these relationships as households move along their life cycle stages and the individual’s commitments and constraints change and these are depicted in the activity-based model; and
- (c) explicit consideration of constraints by the spatial, temporal, and social dimensions of the environment is given. These constraints can be explicit models of timespace prisms (Pendyala 2003) or reflections of these constraints in the form of model parameters and/or rules in a production system format (Arentze and Timmermans 2000).

Input to these models are the typical regional model data of social, economic, and demographic



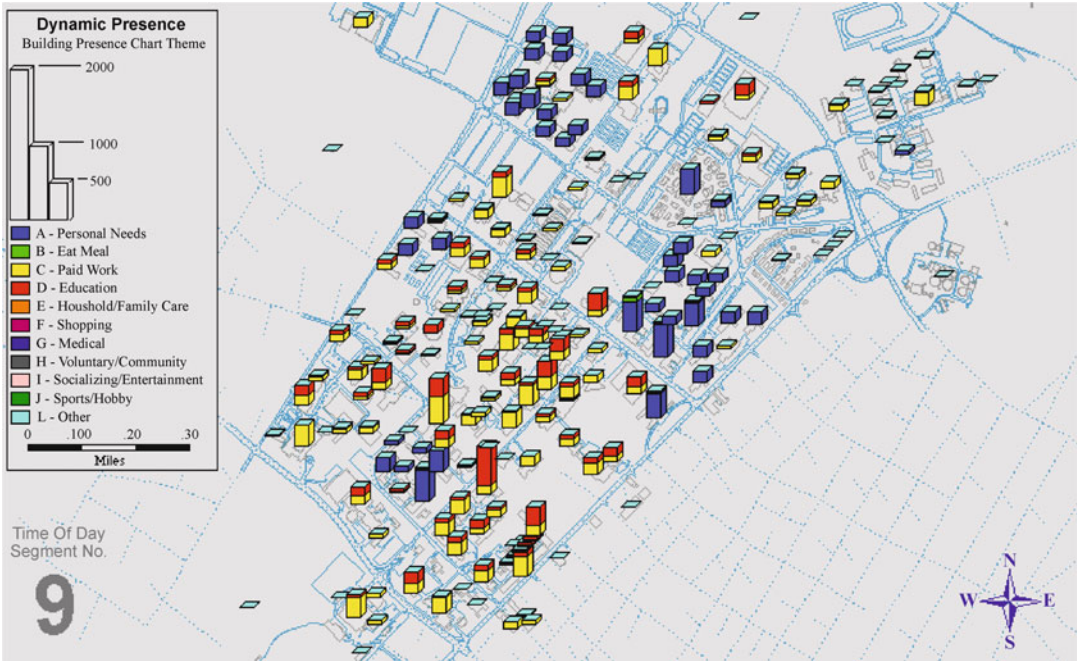
Travel Behavior and Demand Analysis and Prediction, Fig. 2 Time allocation to different activities in a day (Alam 1998). A: Personal Needs (includes sleep), B: Eat

meal, C: Paid work, D: Education, E: Household and family care, F: shopping, G: medical, H: Volunteering/Community, I: Socializing, J: Sports and Hobbies, K: Travel, L: All other

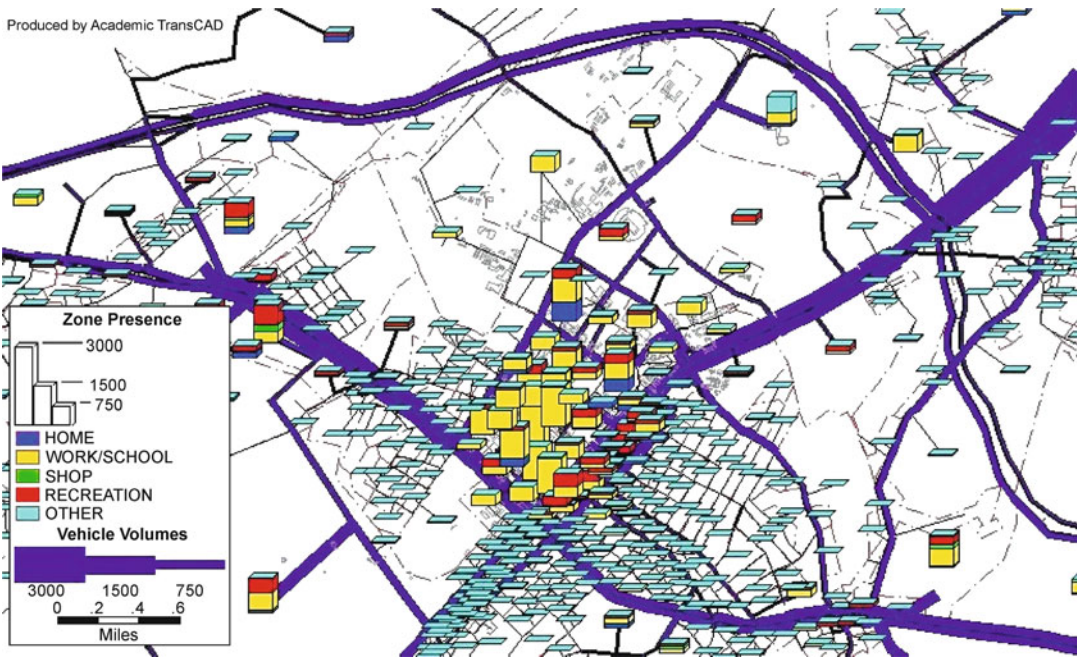
information of potential travelers and land use information to create schedules followed by people in their everyday life. The output are detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity (including travel time, mode used, and so forth). This output is very much like a “day-timer” for each person in a given region. Figure 2 provides an example of time allocation to different activities from an application that collected activity participation data (Alam 1998; Alam and Goulias 1999). It displays time allocation by one segment of the population showing the proportion of persons engaging in each activity by each hour of a day. Figure 3 shows the output from a model that predicts the presence of persons in each building during each hour of a day engaging in each activity type. Combining an activity model with a typical travel demand model produces “volumes” of individuals at specific locations and on the network of a city as shown in Fig. 4 (a more detailed description of this study can be found in

(Goulias et al. 2004; Kuhnau 2001; Kuhnau and Goulias 2003)).

Many planning and modeling applications, however, aim at forecasting. Inherent in forecasting are the time changes in the behavior of individuals and their households and their response to policy actions. At the heart of behavioral change are questions about the process followed in shifting from a given pattern of behavior to another. In addition to measuring change and the relationships among behavioral indicators that change in their values over time, we are also interested in the timing, sequencing, and staging of these changes. Moreover, we are interested in the triggers that may accelerate desirable or delay undesirable changes and the identification of social and demographic segments that may follow one time path versus another in systematic patterns. Knowledge about all this is required to design policies but it is also required to design better forecasting tools. Developments in exploring behavioral dynamics and advancing models



Travel Behavior and Demand Analysis and Prediction, Fig. 3 Persons and activities assigned to buildings (Alam 1998)



Travel Behavior and Demand Analysis and Prediction, Fig. 4 Persons and activities assigned to buildings and travel to the network (Goulias et al. 2004)

for them have progressed in a few arenas. First, in the data collection arena with panel surveys, repeated observation of the same persons over time that are now giving us a considerable history in developing new ideas about data collection but also about data analysis (Golob et al. 1997; Goulias and Kim 2003) and interactive and laboratory data collection techniques (Doherty 2003) that allow a more in-depth examination of behavioral processes. The second arena is in the development of microeconomic dynamic formulations for travel behavior that challenge conventional assumptions and offer alternative formulations (Kitamura 2000). The third arena, is in the behavior from a developmental viewpoint as a single stochastic process, a staged development process (Goulias 1999), or as the outcome from multiple processes operating at different levels (Goulias 2002). Experimentation with new theories from psychology emphasizing development dynamics is a potential fourth area that is just beginning to emerge (Goulias 2003). Behavioral dynamics are also examined using more comprehensive analyzes (Goulias et al. 2007) and models (Ramadurai and Srinivasan 2006).

These models focus more on the paths of persons in space and time within a somewhat short time horizon such a day, a week, or maybe a month. The consideration of behavioral dynamics has expanded the temporal horizons to a few years. However, regional simulation models are very often designed for long range plans spanning 25 years or even longer time horizons. Within these longer horizons, changes in the spatial distribution of activity locations and residences (land use) are substantial, changes in the demographic composition and spatial distribution of demographic segments are also substantial, and changes in travel patterns, transport facilities, and quality of service offered can be extreme. Past approaches in modeling and simulating the relationship among land use, demographics, and travel in a region attempted to disengage travel from the other two treating them as mutually exogenous. As interactions among them became more interesting and pressing, due to urban sprawl and suburban congestion, increasing attention was paid to their complex interdependencies.

This led to a variety of attempts to develop “integrated model systems” that enable the study of scenarios of change and mutual influence between land use and travel. An earlier review of these models with heavy emphasis on discrete choice models can be found in Anas (1982). Miller (2003) and Waddell and Ulfarsson (2003) 20 years later provide two comprehensive reviews of models that have integrated many aspects in the interdependent triad of demographics-travel-land use models. Both reviews trace the history of some of the most notable developments and both link these models to the activity based approach above. Both reviews also agree that a microeconomic and/or macroeconomic approach to modeling land and transportation interactions are not sufficient and more detailed simulation of the individuals and their organizations “acting” in an time-space domain need to be simulated in order to obtain the required output for informed decision making. They also introduce the idea of simulating interactive agents in a dynamic environment of other agents (multi-agent simulation). The vast literature is reviewed by Timmermans (2003) and Miller (2006), from different viewpoints about progress made until now. However, they both agree that progress is rapidly made and that integration of land use and transportation models needs to move forward. Creation of integrated systems is further complicated by the emergence of an entire infrastructural system as another layer of human activity – telecommunication. Today telecommunication and transportation relationships are mostly absent from regional simulation planning and modeling as well from the most advanced land use and transportation integrated models. Considerable research findings, however, have been accumulating since the 1970s (Golob 2001; Goulias et al. 2003; JHK and Associates et al. 1996; Krizek and Johnson 2003; Mahmassani and Jou 1998; Marker and Goulias 2000; Mokhtarian 1990; Patten and Goulias 2001; Patten et al. 2003; Salomon 1986; Weiland and Purser 2000). Another type of technologies (named enabling herein) helped us move modeling and simulation further.

A few of the most important technologies are *stochastic simulation*, *production systems*,

geographic information systems, interactive and technology-aided data collection approaches, and more flexible data analysis techniques.

Stochastic microsimulation, as intended here, is an evolutionary engine software that is used to replicate the relationships among social, economic, and demographic factors with land use, time use, and travel by people. As discussed above the causal links among these groups of entities are extremely complex, non-linear, and in many instances unknown or incompletely specified. This is the reason that no closed form solution can be created for such a forecasting model system. An evolutionary engine, then, provides a realistic representation of person and household life histories (e.g., birth, death, marriages, divorces, birth of children, etc.), spatio-temporal activity opportunity evolution, and a variety of models that account for uncertainties in data, models, and behavioral variation (see Goulias 2002; Miller 2003), for overviews and (Sundararajan and Goulias 2002) for an application).

Production systems were first developed by Newell and Simon (Newell and Simon 1972) to explicitly depict the way humans go about solving problems. These are a series of condition-action (note the parallel with stimulus-response) statements in a sequence. From this viewpoint they are search processes that may never reach an absolute optimum and they replicate (or at least attempt to) human thought and action. Models of this kind are called *computational process models* (CPM) and through the use of IF ... THEN ... rules have made possible the creation of a variety of new models.

Geographic information systems are software systems that can be used to collect, store, analyze, modify, and display large amounts of geographic data. They include layers of data that are able to incorporate relations among the variables in each layer and allow to build relationships in data across layers. One can visualize a GIS as a live map that can display almost any kind of spatio-temporal information. Maps have been used by transportation planners and engineers for long time and they are a natural interface to use in modeling and simulation. GIS today is moving

beyond this relational database definition and is transforming the entire field into GI Science, which is beyond the scope of this article.

Advanced data collection methods and devices that are technologies that merit a note, although, not strictly developed for modeling. The first is about data collection and particularly data collection using internet technologies to build complex interviews that are interactive and dynamic (Doherty 2003). In the same line of development we also see the use of geographic positioning systems (GPS) that allow one to develop a trace of individual paths in time and space (Doherty et al. 2001; Wolf et al. 2001). Very important development is also the emergence of devices that can record the bulk of environmental data surrounding a person's movement, classify the environment in which the individual moves, and then ask simplified questions (Hato 2006).

Soft computing and non-parametric data analysis is the last innovation mentioned here. In the data analysis we see greater strides in using data mining and artificial intelligence-born techniques to extract travel behavior patterns (Pribyl and Goulias 2003; Teodorovic and Vukadinovic 1998) and advanced and less restrictive statistical methods to discover relationships in travel behavior data (e.g., (Kharoufeh and Goulias 2002)). Soft computing is increasingly finding many applications in activity-based models (see www.imob.uhasselt.be). For a more recent and accessible review see Pribyl (Pribyl 2007).

The Evolving Modeling Paradigm

Policies are dictating to create and test increasingly more sophisticated policy assessment instruments that account for direct and indirect effects of behavior, procedures for behavioral change, and to provide finer resolution in the four dimensions of geographic space, time, social space, and jurisdictions. Dynamic planning is also stressing the need to examine trends, cycles, and the inversion of time progression to develop paths from the future visions to today's actions. New model developments are also becoming increasingly urgent. Although, tremendous progress has been

observed in the past 20 years, development requires a faster pace to create new policy tools. These policy tools need to disentangle the actions of persons under different policy actions and the impact of policy actions on aggregates to identify conflicts and resolutions. Supporting all this is a rich collection of decision paradigms that are already used and a few new ideas are starting to migrate to practice as illustrated below.

Early models incorporating activity-based behavioral processes into applications were published in the late 1970s and early 1980s as proof-of-concept and experimental applications. Following Hagerstrand's time-geography approach, PESASP (Lenntorp 1976) is one of the first models to operationally show the use of a time-space prism and to account for the relationship among activities. The Cullen and Godson (1975) study was also the first comprehensive treatment of activities that brought different research findings together. In parallel, models were developed that were utility-maximizing models such as Adler and Ben-Akiva model (Adler and Ben-Akiva 1979) and much later the Kawakami and Isobe model (Kawakami and Isobe 1989). Following these studies, BSP (Huigen 1986) and Computational Algorithms for Rescheduling Lists of Activities – CARLA (Jones et al. 1983) also use the activities within a time-space prism paradigm and define the foundations of data collection for activity-based approaches.

After this period of experimentation three streams in model development emerged. The first is in deriving representative activity patterns (RAPs) and then using regression techniques to correlate RAPs to person and household social and demographic data and then forecasting. The second development refines the methods used to simulate persons and adds to the forecasting repertoire other forecasting tasks via *micro-simulation*. The third is a movement that expands the envelope to include cognition and explicit representation of mental processes through CPMs.

The Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions (STARCHILD – Recker and McNally

(Recker et al. 1986a, b) derived RAPs, employed a utility-based model and incorporated constraints. It is considered a fundamental transition development from research to practical application of an activity-based approach and it is still the foundation of models that first derive representative patterns and then forecast travel behavior. The more recent SIMAP (Kulkarni and McNally 2001) is a direct derivation of STARCHILD. In this line of development, Ma (1997) created a model system that combined long term activity patterns (Long-term activity and travel planning – LATP) with a within-a-day activity scheduling and simulation (Daily Activity and Travel Scheduling – DATS) incorporating day-to-day variation and history dependence. Her model system produced very accurate forecasts. However it required panel survey data (the repeated observation of the same persons and households over time) that are rarely collected. In the LATP/DATS system longitudinal statistical models are extracted from longitudinal records and they capture important aspects of behavioral dynamics such as habit persistence, day-to-day switching behaviors, and account for observed and unobserved heterogeneity contributed by the person, the household, the area of residence, and the area of workplace.

One of the first models to include a micro-simulation in its paradigm is ORIENT (Sparmann 1980). This methodology suitably refined was demonstrated in a countrywide model for the Netherlands developed between 1989 and 1991 and named the Microanalytic Integrated Demographic Accounting System (MIDAS – Goulias and Kitamura (1992, 1997)). MIDAS integrates demographic microsimulation, with dynamic car ownership models and a comprehensive suite of travel behavior equations. A cross-sectional version of MIDAS using data from the United States was also developed by Chung and Goulias (1997). MIDAS-USA simulates the evolution of households along with car ownership and travel behavior for Centre County, PA, and it is linked to a model to assign fees for development using GIS. A more ambitious development is the Activity Mobility Simulator – AMOS – by Kitamura et al. (1996), which defines

a few RAPs as templates. Then, uses a neural network to identify choices and a satisficing rule to simulate schedule changes due to policies. While MIDAS is a strictly longitudinal process econometric model progressing 1 year at a time, AMOS is constraint-based model designed for much finer temporal resolution. DEMOS, developed by Sundararajan and Goulias (2002), is another MIDAS derivative. DEMOS is an object-oriented environment designed to simulate the evolution of people and their households using a variety of external data with the core models based on the Puget Sound Transportation Panel. It also simulates activity participation, travel, and telecommunication market penetration using a few representative patterns that were derived in Ma's LATP/DATS supplemented by telecommunication and travel behavior models.

SCHEDULER (Gärling et al. (1989) is the first CPM that adds a psychometric cognitive implementation based on the Hayes-Roth and Hayes-Roth (1979) planning model. In SCHEDULER, activities, selected from the long term calendar that represents a person's long term memory, comprise a schedule that is "mentally executed." Models start to combine CPM, microsimulation, and data derived behavioral patterns with random utility models to fill different modeling needs. The Simulation Model of Activity Scheduling Heuristics (SMASH – Ettema et al. (1996)) is a CPM and econometric utility-based hybrid model that focuses on the pre-trip planning process predicting sequences of activities. In parallel, COMRADE (Ettema et al. 1995), uses competing risk hazard models for activity scheduling and incorporates duration models in the system. The Model of Action Space in Time Intervals and Clusters (MASTIC – Dijst and Vidakovic (1997)), identifies clusters in the action space to perform and schedule activities. Time-space prisms are also the foundation of the Prism-Constrained Activity-Travel Simulator (PCATS – Kitamura (1997), Kitamura and Fujii (1998)), which is also a utility-based model. A direct operational derivative of SCHEDULER (Gärling et al. 1994) was developed by Kwan, in her 1994 dissertation (Kwan 1994, 1997), and named GIS-Interfaced Computational-process modeling for Activity

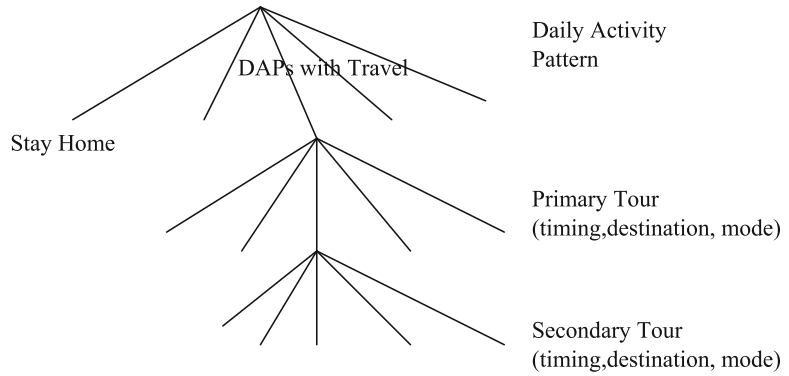
Scheduling (GISICAS). It is a simplified CPM, that uses time-space constraints and GIS to incorporate spatial information into a behavioral model to create individual schedules, starting with activities at higher levels of priority. Other models also attempt to recreate personal schedules such as Vause's model (Vause 1997), a CPM that creates a restricted choice set for creating activity patterns, a model by Ettema (Ettema et al. 1997), and VISEM (Fellendorf et al. 1997), a data-driven model that is a part of PTV Vision, an urban and regional transportation planning system, that creates daily activity patterns for behaviorally homogeneous groups within the population. Stopher et al. (1996) also proposed the Simulation Model for Activity Resources and Travel (SMART) using a time geography framework and a taxonomy of activities in a GIS environment. All these use observed patterns to derive behavioral models. In contrast, Recker (1995), developed the Household Activity Pattern Problem (HAPP) as a normative model based on the pick up and delivery time window problem to be used as a yardstick model testing optimal behavioral hypotheses.

The model framework that impacted practice the most in the United States is the Daily Activity Schedule model by Ben-Akiva et al. in (1996). This model, was used to create the Portland Daily Activity Schedule Model (Bowman et al. 1998), advocating modeling lifestyle and mobility decisions on a scale of years. These influence daily activity schedules, which are comprised of primary and secondary tours constrained in time and space. It contains two key elements that simplify activity-based model development and takes advantage of the research surge in developing more general discrete choice models. A similar simplification using conditional probabilities was also developed for Los Angeles by Kitamura et al. (1997).

Figure 5 shows this hierarchy of decisions and the scheme used to convert the daily pattern into a system of discrete choices. This framework was used to design new models for the regions around San Francisco, New York, Columbus, Denver, Atlanta, and Sacramento (Bradley and Bowman 2006).

Travel Behavior and Demand Analysis and Prediction, Fig. 5

The Bowman and Ben-Akiva daily activity model formulation (Bowman et al. 1998)



Arentze and Timmermans (2000) designed the most complete CPM named ALBATROSS, which is a multi-agent simulation and predicts the time, location, duration, activity companionship, and travel modes subjecting everything to spatio-temporal, institutional, and household constraints. The theoretical underpinnings of this model are by far wider and all encompassing than any other activity-based model. However, it does not simulate route choice and does not produce data suitable for traffic assignment algorithms. Development of the third version of ALBATROSS is currently underway (Henson et al. 2006). This model is also representative of raising the ambitions of travel modelers. The Alam Penn State Emergency Management model (Alam-PSEM, Alam and Goulias (1999)) is a building-by-building simulation of activity participation and presence at specific locations of a university campus for each hour of a typical day. In parallel Bhat and his co-workers (Bhat 2001; Bhat and Singh 2000) developed the Comprehensive Activity-Travel Generation System for Workers (CATGW), which is a series of econometric models that replicate a commuter's evening mode choices, number of evening commute stops, and the number of stops after arriving home. The models developed by Bhat and colleagues are characterized by the use of hazard/duration regression models that were specifically developed for activity-based approaches and are by far more flexible than other regression methods. Another econometric model, the Conjoint-Based Model to Predict Regional Activity Patterns (COBRA), developed by Wang and Timmermans in (Wang and Timmermans 2000), generates general patterns of stops for specific activities using a

conjoint-based model with stated preference data instead of travel or activity diary data. The Wen and Koppelman model (Wen and Koppelman 2000) utilizes three layers of decisions that are influenced by exogenous variables to generate activity patterns.

All these models point to new directions such as spatial choice needs to be dealt in more detail (Alam and Goulias 1999), activity choice and duration need to be dealt in a way that recognizes satiation in activity participation (e.g., in the duration models of Bhat (2001)), sooner or later we will need to account for unobserved patterns and lack of experimental data (e.g., using conjoint experiments Wang and Timmermans (2000)), and relations within the household need to also receive attention and be inserted in the model hierarchy (Wen and Koppelman 2000).

Spatial aspects of model development were considered in the CentreSIM regional model (Goulias et al. 2004; Kuhnu and Goulias 2002, 2003) that uses time-of-day activity and travel data for different market segments to predict hour-by-hour presence at locations and travel among zones. In 2004, as a part of the Longitudinal Integrated Forecasting Environment (LIFE) framework (Goulias 2001), Pribyl and Goulias (2005) developed CentreSIM (medoid simulation) to derive a few representative patterns and simulate daily schedules accounting explicitly for within-household interactions for entire daily patterns. In the Netherlands, PATRICIA (Predicting Activity-Travel Interdependencies with a Suite of ChoiceBased, Interlinked Analyses), was developed by Borgers et al. (2002) to help assess the

performance of ALBATROSS. PATRICIA is a suite of linked models that incorporates an expanded set of activity choices, based on 63 distinct patterns, and activity destinations and describes activity transport modes and sequences. AURORA (Joh et al. 2004; Timmermans et al. 2001), which is a complementary model to ALBATROSS, is a utility-based system that models the dynamics of activity scheduling and rescheduling decisions as a function of many choice facets. AURORA is for short-term adaptation and rescheduling using just a few critical parameters. The model has since been expanded to include many new facets (Henson et al. 2006). A much simpler model is PETRA (Fosgerau 2002) that allows the model to work with a small number of daily travel patterns with some statistical advantages (see also Henson et al. 2006). Microsimulation software experienced another push forward by the development of a multi-million investment in TRansportation ANalysis SIMulation System. This model system was developed in the decade 1995–2005 and one of its versions is now available via a NASA open source license from TMIP at <http://tmip.fhwa.dot.gov/transims/>. TRANSIMS is a survey data-driven cellular automata microsimulation and was developed by a team at Los Alamos National Laboratory (2003). It was one of the first simulation packages to contain models that create a synthetic population, generate activity plans for individuals using directly observed data in travel surveys, formulate routes on a network based on these, and execute activity plans. Microsimulation models also evolved in the interface between land use and travel behavior. The Integrated Land Use, Transportation and Environment (ILUTE) model (Salvini and Miller 2003) model is designed to simulate the evolution of people and their activity patterns, transportation networks, houses, commercial buildings, the economy, and the job market over time. Within this vision, Miller and Roorda (2003), developed the Toronto Area Scheduling model for Household Agents (TASHA) that uses *projects* to organize activity episodes into schedules of persons. Schedules for members in a household are simultaneously generated to allow for joint activities. Both ILUTE

and TASHA utilize CPMs and econometric utility based paradigms.

Another microsimulation that uses econometric models to simulate daily activity travel patterns for an individual, is the Comprehensive Econometric Microsimulator for Daily Activity-travel Patterns (CEMDAP) model (Bhat et al. 2003) that is based on land use, socio-demographic, activity system, and level-of-service (LOS) attributes. Key distinctive element of CEMDAP is its reliance on hazard-based regression models to account for the continuous nature time of activity duration. It includes population synthesis as well as the activity-pattern generation and scheduling of children, which is missing from many other simulators. Another model that utilizes constraints is the Florida Activity Mobility Simulator (FAMOS) (Pendyala et al. 2005). FAMOS encompasses two modules, the Household Attributes Generation System (HAGS) and PCATS. Together, they comprise a system for modeling the activity patterns of individuals in Florida. The output is a series of activity-travel records. FAMOS is currently being further enhanced to include intra-household interactions and capture task allocation behavior among household members. Most recently, Ettema et al. (2006) developed PUMA (Predicting Urbanization with Multi-Agents), a full-fledged multi-agent system of urban processes that represents land use changes in a behaviorally realistic way. These processes include the evolution of population, businesses, and land use as well as daily activity and travel patterns of people. To simulate activity-travel patterns, an updated version of AURORA by Arentze et al. (2006) will be created and also in the model FEATHERS (Forecasting Evolutionary Activity-Travel of Household and their Environmental Repercussions) to simulate activity-level scheduling decisions, within-a-day rescheduling, and learning processes in high resolutions of time and space. Developed as a complement to ALBATROSS, FEATHERS is econometric utility-based microsimulation that utilizes constraints that focuses on the short-term dynamics of activity-travel patterns. Members from this same Dutch team also developed MERLIN (Van Middelkoop et al. 2004) and RAMBLAS (Veldhuisen et al. 2000).

Microsimulations have continued to gain in popularity in the activity-based modeling universe as they move from research applications to practice. Besides the Portland Daily Activity Schedule Model mentioned previously, New York's "Best Practice" Model (2002) and the Mid-Ohio Regional Planning Commission (MORPC) Model (Vovsha et al. 2003), both developed by Vovsha et al., and the San Francisco model (Jonnalagadda et al. 2001) are currently being utilized by their respective MPO. The San Francisco model is currently being updated to implement enhanced destination choice models and being recalibrated using more recent household and census data. Four other models for Atlanta, Sacramento, the San Francisco Bay Area, and Denver are currently in various stages of implementation (Bradley and Bowman 2006).

Although many past activity-based models have undefined or large time resolutions, STARCHILD already in mid-1980s used 15-min temporal resolution. The most recent models, however, go even further to simulate activities at small time intervals such as 5 min (TASHA) and 10 min intervals (SIMAP), minute by minute (MASTIC, CentreSIM, MASTIC, GISICAS, and RAMBLAS), and second-by-second (TRANSIMS-LANL, ALBATROSS, AURORA, CATGW, CEMDAP, FAMOS, and FEATHERS). Many applications, however, operate with large resolutions of 1 h and they are implemented with a target of 30 min to 1 h (Bradley and Bowman 2006). Spatial resolution of the models is still dominated by the zonal level. ALBATROSS and MORPC both can operate at the subzone level. Alam-PSEM, AURORA, CEMDAP, FEATHERS, GISICAS, ILUTE, PUMA, SIMAP, SMASH, and TRANSIMS-LANL utilize data at essentially the building or point level. Only two applications have spatial resolutions below the zonal level (Denver model that contains a two-stage destination locator to predict the address within a zone and the Sacramento model that operates at the parcel level). Cognitive theories (models of knowledge and memory as well as behavioral process for planning activities) were used only in SCHEDULER and based on that in ALBATROSS and FEATHERS. Behavior is most

often incorporated as intra-household interaction in ALBATROSS, CEMDAP, FAMOS, FEATHERS, ILUTE/TASHA, and CentreSIM as well as some of the applications in regions such as MORPC.

Examples of Mathematical Models

In this section additional details of two examples of mathematical models for activity and travel behavior analysis are offered. Both examples aim at incorporating human interaction in time allocation models and they are multilevel regression models (based on Goulias (2002)) and group decision making utility maximization models (based on Zhang et al. (2005)).

Multilevel Regression Models

These regression models are known by different names in different fields of research such as random coefficient models ((Greene 1997) and p. 669 in (Longford 1993)), multilevel models (Goldstein 1995), mixed models (Searle et al. 1992), and hierarchical linear models (Bryk and Raudenbush 1992). They describe the contextual nature of the data and/or the way of accounting for dependent variable variation from multiple sources. Key advantages of these models are: explicit recognition in model formulation of the hierarchical, multiple level and nested structure of the data we analyze, and model specification using three groups of regression components in the same regression model. The first group assumes constant sensitivity to explanatory variables among the units of analysis representing the mean effect of an explanatory variable on the dependent variable. The second group assumes a random deviation around this mean and the third group is the usual random error term(s) of the regression equation. When compared to traditional regression models, which contain only one level, multilevel models do not underestimate the standard errors of coefficient estimates avoiding overstatements about the statistical significance of policy variables (e.g., we do not exaggerate the effect of taxation on car ownership or the effect of

time and cost on route choice). A system of multi-level regression models can be written as follows.

$$Y_{ij}^q = \alpha_{ij}^q + \beta_k^q X_{ij} + \gamma_m^q Z_{ij} \quad (1)$$

$$\alpha_{ij}^q = \gamma_0^q + v_j^q + u_{ij}^q + \varepsilon_{ij}^q, \quad \text{where} \quad (2)$$

$$q = 1, \dots, Q,$$

$$\beta_{k1}^q = \gamma_{k1}^q + u_{k1j}^q, \quad \beta_{k2}^q = \gamma_{k2}^q + v_{k2j}^q. \quad (3)$$

Equation (1), represents Q equations that are one for each Y_{ij}^q variable that we want to explain and use in travel demand forecasting. They can be the amount of time dedicated to activities and travel or distances to specific destinations or even attributes of routes considered by trip makers. The index t represents the time at which an observation was made for a person i from within a household j (with $t = 1, 2, 3, \dots, T$, $i = 1, 2, \dots$, number of people in household j , $j = 1, 2, \dots$, number of households in sample). In this way we can identify change from one time point to another by an individual and study the relationships among individuals within social units (e.g., households, associations, neighborhoods and so forth).

The time points can be the same for all individuals or they may vary depending on the data collection procedures and willingness of respondents to provide information. Equation 1 is called the level 1 model because it is written at the level of the time point (observation occasion). The first term in the right hand side of Eq. (1) is a random intercept, α , given by Eq. (2). This component has specific meaning. For example, α_{ij}^q is the mean value of person i in household j at time t for variable q . The term ε_{ij}^q is a random temporal variation (also called within person variation) and it is the deviation of time expenditure around γ_0^q . The term u_{ij}^q is a random person to person variation (also called within household variation) and it is also a deviation of around γ_0^q . The term v_j^q is a random household to household variation and it is also a deviation around γ_0^q . These are also called random error components and they are usually assumed normally distributed with $E(\varepsilon) = E(u) = E(v) = 0$, with $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$, $\text{Var}(u) = \sigma_u^2$, and $\text{Var}(v) = \sigma_v^2$ to be estimated. It is worth noting

that the system of equations represented by Eq. (1) contain a set of gamma coefficients (associated with a matrix Z representing explanatory variables) that are defined in a similar way as in typical regression models. The β s, however, that multiply the matrix X are defined as random with a mean and a variation around the means γ s. This variation can be due to the temporal, personal, and/or household levels. In this way, we can define a variety of equations at each of these levels to represent heterogeneous behavior that is either due to temporal fluctuations, personal variation, or household variation. Equation (3) differentiates between β s that vary within individuals and those that vary within households. In this way, at each level we have a level-specific variance-covariance matrix of all the random terms (ε s, u s, v s). The significance of the elements in each of these three matrices can be tested using goodness-of-fit measures based on the deviance, which is the difference in the $-2\text{Log}(\text{likelihood})$ at convergence between two nested (in terms of specification) models. In addition, the γ s can also be tested if they are significantly different than zero using a t -test. The γ s in Eq. (1) are called the *fixed effects* and the remaining terms are called the *random effects* at each of the three levels in the hierarchy. Estimation of all the fixed and random parameters can be accomplished either by Full Information Maximum Likelihood, FIML, applied to Y directly or applied to the least-squares residuals, called Restricted Maximum Likelihood-REML that can be used in tandem with a generalized least squares approach. Longford (1993), Bryk and Raudensbush (1992; Goldstein 1995) provide a comprehensive review of estimation techniques, their performance assessment, and detailed algorithms.

Household Utility Models

The second example is also representative of a movement toward more detailed consideration of within household decision making dynamics. Although the model was specified by Zhang et al. (2005) for time allocation to shared (j) and non-shared activities (s), it is a potentially useful model for other trip making decisions. Each person in a household is assumed to form two utility functions. One utility is for the shared activities (i.e., engagement in activities with other

household members) and non-shared activities. These utility functions are given by Eq. (4) (shared activity) and 5 (non-shared activity).

$$u_{is} = \exp \left(\left(\alpha_s + \sum_k \beta_{sk} x_{isk} \right) \ln \left(\sum_m \kappa_{sm} \tau_{ism} \right) + \varepsilon_{is} \right) \ln(t_{is}) \quad (4)$$

$$u_{ij} = \exp \left(\left(\alpha_j + \sum_k \beta_{jk} x_{ijk} \right) \ln \left(\sum_m \kappa_{jm} \tau_{ijm} \right) + \varepsilon_{ij} \right) \ln(t_{ij}), \quad (5)$$

where:

- α_s is the constant term for each shared activity s .
- x_{isk} is the k th explanatory variable (and/or attribute) of household member i for shared activity s .
- β_{sk} is the parameter associated with the k th attribute of the shared activity.
- τ_{ism} is the travel time by mode m for each activity s by person i .
- κ_{sm} is the parameter associated with travel time by mode m .
- t_{is} is a random error term of the shared activity s by person i .
- α_j is the constant term for each non-shared activity.
- x_{ijk} is the k th explanatory variable (attribute) of household member i for non-shared activity j .
- β_{jk} is the parameter associated with the k th attribute of non-shared activity.
- τ_{ijm} is the travel time by mode m for each activity s by person i .
- κ_{jm} is the parameter associated with travel time by mode m .
- ε_{ij} is a random error term of the non-shared activity j by person i .
- t_{ij} is the amount of time dedicated to activity j by person i .

The overall utility of activity participation and travel for each person i under the assumption of a multi-linear utility is given by Eq. (6).

$$u_i = \sum_{j=1}^{J+S} r_{ij} u_{ij} + \sum_{j=1}^{J+S} \sum_{j'>j} \delta_i r_{ij} r_{ij'} u_{ij} u_{ij'} \quad (6)$$

Where

- u_{ij} is the utility of activity j for person i .
- r_{ij} is the relative interest of person i for activity j .
- δ_i is parameter of activity dependency for member i .
- $J + S$ is the number of non-shared and shared activities for a person within the unit of time under consideration.

In a similar way the household utility function is a multi-linear combination of the individual utilities in Eq. (7).

$$\text{HUF} = \sum_{i=1}^n w_i u_i + \lambda \sum_{i=1}^n \sum_{i'>i} (w_i w_{i'} u_i u_{i'}) \quad (7)$$

where,

HUF is the household utility combining the utilities of all household members n .

u_i is the utility of household member i .

w_i is the relative influence of each household member i .

λ is a parameter of within household interaction.

Under the assumption of maximizing HUF it is possible to create a Lagrangian function that accounts for constraints (i.e., total amount of time available, signs of parameters and so forth) and through a maximization solution derive equations that can be used to estimate the unknown parameters in Eqs. 4, 5, 6, and 7 (details are provided in Zhang et al. (2005) for time allocation). It is worth noting that Zhang et al. (2005), derived two alternate model systems by changing the utility functions to represent different intra-household bargaining models (for a detailed review see Bergstrom 1995). Then through a linearization process they developed a system of linear equations and estimated the parameters using a multiple equations econometric approach (the Seemingly Unrelated Regression Estimation

Greene 1997) that is a simplifying alternative to the multilevel models described earlier in this section. A more general review of this type of model formulation is also provided by Timmermans (Timmermans 2006).

Summary

Similarities and differences among the implemented modeling ideas are:

- A hierarchy of decisions by households is assumed that identifies longer term choices determining the shorter term choices. In this way different blocks of variables can be identified and their mutual correlation used to derive equations that are used in forecasting.
 - Anchor points (Home location – work location school location) are inserted in the first choice level and they define the overall spatial structure of activity scheduling.
 - Out-of-home activity purposes include work, school, shopping, meals, personal business, recreation, and escort. These expand the original home-based and nonhome based purposes in travel behavior and the three activity types in home economics (labor for pay, labor at home, and leisure).
 - In-home activities are explicitly modeled or allowed to enter the model structure as a “stay-at-home” choice with some models allowing for activity choice at home (work, maintenance and discretionary). In this way home can be reflected in the models.
 - Stop frequencies and activities at stops are modeled at the day pattern and tour levels to distinguish between activities and trips that can be rescheduled with little additional efforts versus the activities and trips that cannot be rescheduled (e.g., school trips).
 - Modes and destinations are modeled together. In this way the mutual influence – sequential and/or simultaneous relationships can be reflected in the model structure.
 - Time is included in a few instances in activity-based models. For example departure time for trips and tour time of day choice are modeled explicitly. Model time periods are anywhere between 30 min and second-by second and time windows are used to account for scheduling. This modeling component allows to incorporate time-of-day in the modeling suites. It also allows to identify windows of activity and travel opportunities. The presence of departure time also enables models to trip matrices for any desired periods in a day. In fact, output of time periods depends on route choice and traffic assignment needs and can be adjusted almost at will.
 - Human interaction, although limited for now to the within-household interaction, is incorporated by relating the day pattern of one person to the day patterns of other persons within a household, their joint activities and trip making are explicitly modeled (joint recreation, escort trips), and allocation of activity-roles are also modeled.
 - Spatial aspects of a region are accounted for using methods that produce spatially distributed synthetic populations using as external control totals averages and relative frequencies of population characteristics.
 - Accessibility measures are used to capture spatial interaction among activity locations and the level of service offered by the transportation systems. These are also the indicators used to account for feedback among the lower level in the hierarchy decisions (e.g., activity location choices, routes followed, congestion) and the higher level such as residence location choice.
 - Spatial resolution is heavily dependent on data availability and it reached already the level of a parcel and/or building at its most disaggregate level. Outputs of models are then aggregated to whatever level is required by traffic assignment, mode specific studies (nonmotorized and/or transit) and reporting needs and requirements.
- Overall, the plethora of advances includes: a) models and experiments to create computerized virtual worlds and synthetic schedules at the most elementary level of decision making using micro-simulation and computational process models; b)

data collection methods and new methods to collect extreme details about behavior and to estimate, validate, and verify models using advanced hardware, software, and data analysis techniques; and c) integration of models from different domains to reflect additional interdependencies such as land use and telecommunications.

Future Directions

Much more work remains to be done in order to develop models that can answer more complex questions in policy analysis and for this reason a few steps are outlined here. In policy and program evaluation, transportation analysis appears to be narrowly applied to only one method of assessment that does not follow the ideal of a randomized controlled trial and does not explicitly define what experimental setting we are using for our assessments. Unfortunately this weakens our findings about policy analysis and planning activities. Although we have many laboratory experiments that were done for intelligent transportation systems we lack studies and guidelines to develop experimental and quasi-experimental procedures to guide us in policy development and large scale data collection.

In addition, many issues remain unresolved in the areas of coordination among scale in time and space and related issues. In addition very little is known about model sensitivity and data error tolerance and their mapping to strategy evaluations. This is partially due to the lack of tools that are able to make these assessments but also due to lack of scrutiny of these issues and their implications on impact assessment.

Regarding strategic planning and evaluation, we also lack models designed to be used in scenario building exercises such as backcasting and related assessments. The models about change are usually defined for forecasting and simple time inversion may not work to make them usable in backcasting. This area does not have the long tradition of modeling and simulation to help us develop suitable models. Should more attention be paid to this aspect? Is there room for a combination of techniques including qualitative

research methods? What is the interface between this aspect and the experimental methods questions in program evaluation?

In the new research and technology area, since we are dealing with the behavior of persons, it is unavoidable to consider perceptions of time and space. What role should perceptions of time and space (Golledge and Gärling 2004) play in behavioral models and what is the most appropriate use of these perceptions? The multiple dimensions of time such as tempo, duration, and clock time are neglected in behavioral models – is there a role for them in behavioral models?

Human interaction is considered important and is receiving attention in more recent research Golob and McNally (1997), Chandrasekharan and Goulias (1999), Simma and Axhausen (2001), Gliebe and Koppelman (2002), Goulias and Kim (2005), Zhang et al. (2005), but only partially accounted for in applications as illustrated by Vovsha and Petersen (2005). Future applications will increasingly pay attention to motivations for human interactions and the nature of these interactions within households and in a wider social network context.

Bibliography

- Adler T, Ben-Akiva M (1979) A theoretical and empirical model of trip chaining behavior. *Transp Res B* 13:243–257
- Alam BS (1998) Dynamic emergency evacuation management system using GIS and spatio-temporal models of behavior. MS Thesis, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
- Alam BS, Goulias KG (1999) Dynamic emergency evacuation management system using GIS and spatio-temporal models of behavior. *Transp Res Rec* 1660:92–99
- Anas A (1982) Residential location markets and urban transportation: economic theory, econometrics and policy analysis with discrete choice models. Academic, New York
- Arentze T, Timmermans H (2000) ALBATROSS – a learning based transportation oriented simulation system. European Institute of Retailing and Services Studies (EIRASS), Technical University of Eindhoven, Eindhoven
- Arentze T, Timmermans H, Janssens D, Wets G (2006) Modeling short-term dynamics in activity-travel

- patterns: from aurora to feathers. Presented at the innovations in travel modeling conference, Austin, 21–23 May 2006
- Avineri E, Prashker Y (2003) Sensitivity to uncertainty: the need for a paradigm shift. CD-TRB ROM proceedings, paper presented at the 82nd annual transportation research board meeting, 12–16 January 2003, Washington, DC
- Becker GS (1976) The economic approach to human behavior. The University of Chicago Press, Chicago
- Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand. MIT Press, Cambridge
- Ben-Akiva ME, Morikawa T (1989) Estimation of mode switching models from revealed preferences and stated intentions. Paper presented at the international conference on dynamic travel behavior at Kyoto University Hall, Kyoto
- Ben-Akiva M, Bowman JL, Gopinath D (1996) Travel demand model system for the information era. *Transportation* 23:241–266
- Ben-Akiva ME, Walker J, Bernardino AT, Gopinath DA, Morikawa T, Polydoropoulou A (2002) Integration of choice and latent variable models. In: Mahmassani HS (ed) *In perceptual motion: travel behavior research opportunities and application challenges*. Pergamon, Amsterdam
- Bergstrom TC (1995) A survey of theories of the family. Department of Economics, University of California Santa Barbara, Paper 1995D. <http://repositories.cdlib.org/ucsbecon/bergstrom/1995D/>
- Bhat CR (2000) Flexible model structures for discrete choice analysis. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 71–89
- Bhat C (2001) A comprehensive and operational analysis framework for generating the daily activity-travel pattern of workers. Paper presented at the 78th annual meeting of the transportation research board, Washington, DC, 10–14 January 2001
- Bhat CR (2003) Random utility-based discrete choice models for travel demand analysis. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 10-1–10-30
- Bhat CR, Koppelman F (1999) A retrospective and prospective survey of time-use research. *Transportation* 26(2):119–139
- Bhat CR, Singh SK (2000) A comprehensive daily activity-travel generation model system for workers. *Transp Res A* 34(1):1–22
- Bhat CR, Guo J, Srinivasan S, Sivakumar A (2003) Center for Transportation Research, Austin
- Bockenholt U (2002) Comparison and choice: analyzing discrete preference data by latent class scaling models. In: Hagenaars JA, McCutcheon AL (eds) *Applied latent class analysis*. Cambridge University Press, Cambridge, pp 163–182
- Borgers AWJ, Hofman F, Timmermans HJP (1997) Activity-based modelling: prospects. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Pergamon, Oxford, pp 339–351
- Borgers AWJ, Timmermans AH, van der Waerden P (2002) Patricia: predicting activity-travel interdependencies with a suite of choice-based, interlinked analysis. *Transp Res Rec* 1807:145–153
- Bowman JL, Bradley M, Shifan Y, Lawton TK, Ben-Akiva M (1998) Demonstration of an activity-based model system for Portland. Paper presented at the 8th world conference on transport research, Antwerp, June 1998
- Bradley M, Bowman J (2006) A summary of design features of activity-based microsimulation models for US MPOs. Conference on innovations in travel demand modeling, Austin, 21–23 May 2006
- Brög W, Erl E (1980) Interactive measurement methods – theoretical bases and practical applications. *Transp Res Rec* 765:1–6
- Bryk AS, Raudenbush SW (1992) *Hierarchical linear models*. Sage, Newberry Park
- Chandrasekharan B, Goulias KG (1999) Exploratory longitudinal analysis of solo and joint trip making in the Puget Sound transportation panel. *Transp Res Rec* 1676:77–85
- Chapin FS Jr (1974) *Human activity patterns in the city: things people do in time and space*. Wiley, New York
- Chung J, Goulias KG (1997) Travel demand forecasting using microsimulation: initial results from a case study in Pennsylvania. *Transp Res Rec* 1607:24–30
- Creighton RL (1970) *Urban transportation planning*. University of Illinois Press, Urbana
- Cullen I, Godson V (1975) Urban networks: the structure of activity patterns. *Progr Plan* 4(1):1–96
- Dijst M, Vidakovic V (1997) Individual action space in the city. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Elsevier Science, New York, pp 117–134
- Dillman DA (2000) *Mail and internet surveys: the tailored design method*, 2nd edn. Wiley, New York
- Doherty S (2003) Interactive methods for activity scheduling processes. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 7-1–7-25
- Doherty ST, Noel N, Lee M-G, Sirois C, Ueno M (2001) Moving beyond observed outcomes: global positioning systems and interactive computer-based travel behavior surveys. Transportation research circular, E-C026, March 2001. Transportation Research Board, Washington DC
- Ettema DF, Timmermans HJP (1997) *Activity-based approaches to travel analysis*. Elsevier Science, New York, p xiii
- Ettema DF, Borgers AWJ, Timmermans HJP (1995) Competing risk hazard model of activity choice, timing, sequencing and duration. *Transp Res Rec* 1439:101–109
- Ettema DF, Borgers AWJ, Timmermans H (1996) SMASH (simulation model of activity scheduling heuristics): some simulations. *Transp Res Rec* 1551:88–94

- Ettema DF, Daly A, de Jong G, Kroes E (1997) Towards an applied activity-based travel demand model. Paper presented at the IATBR conference, Austin, 21–25 September 1997
- Ettema DF, de Jong K, Timmermans H, Bakema A (2006) PUMA: multi-agent modeling of urban systems. 2006 Transportation Research Board CD-ROM
- Fellendorf M, Haupt T, Heidl U, Scherr W (1997) PTV vision: activity based demand forecasting in daily practice. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Elsevier Science, New York, pp 55–72
- Fosgerau M (2002) PETRA – an activity-based approach to travel demand analysis. In: Lundquist L, Mattsson L-G (eds) *National transport models: recent developments and prospects*. Royal Institute of Technology/Springer, Stockholm/Berlin
- Gärling T, Brannas K, Garvill J, Golledge RG, Gopal S, Holm E, Lindberg E (1989) Household activity scheduling. In: *Transport policy, management and technology towards 2001. Selected proceedings of the fifth world conference on transport research*, vol 4. Western Periodicals, Ventura, pp 235–248
- Gärling T, Kwan M, Golledge R (1994) Computational-process modeling of household travel activity scheduling. *Transp Res Part B* 25:355–364
- Gärling T, Laitila T, Westin K (1998) Theoretical foundations of travel choice modeling: an introduction. In: Gärling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam, pp 1–30
- Garrett M, Wachs M (1996) *Transportation planning on trial. The clean air act and travel forecasting*. Sage, Thousand Oaks
- Gliebe JP, Koppelman FS (2002) A model of joint activity participation. *Transportation* 29:49–72
- Goldstein H (1995) *Multilevel statistical models*. Edward Arnold, London/New York
- Golledge RG, Gärling T (2003) Spatial behavior in transportation modeling and planning. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 1–27
- Golledge RG, Gärling T (2004) Cognitive maps and urban travel. In: Hensher D, Button K, Haynes K, Stopher P (eds) *Handbook of transport geography and spatial systems*, vol 5. Elsevier, Amsterdam, pp 501–512
- Golledge RG, Stimpson RJ (1997) *Spatial behavior: a geographic perspective*. Guilford Press, New York
- Golledge RG, Smith TR, Pellegrino JW, Doherty S, Marshall SP (1985) A conceptual model and empirical analysis of children's acquisition of spatial knowledge. *J Environ Psychol* 5(2):125–152
- Golob TF (2001) Travelbehaviour.com: activity approaches to modeling the effects of information technology on personal travel behaviour, in travel behavior research. In: Hensher D (ed) *The leading edge*. Elsevier Science/Pergamon, Kidlington/Oxford, pp 145–184
- Golob TF, McNally M (1997) A model of household interactions in activity participation and the derived demand for travel. *Transp Res B* 31:177–194
- Golob TF, Kitamura R, Long L (eds) (1997) *Panels for transportation planning: methods and applications*. Kluwer, Boston
- Goodman LA (2002) Latent class analysis: the empirical study of latent types, latent variables, and latent structures. In: Hagenaars JA, McCutcheon AL (eds) *Applied latent class analysis*. Cambridge University Press, Cambridge, pp 3–55
- Goulias KG (1999) Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models. *Transp Res B* 33:535–557
- Goulias KG (2001) A longitudinal integrated forecasting environment (LIFE) for activity and travel forecasting. In: Villacampa Y, Brebbia CA, Uso J-L (eds) *Ecosystems and sustainable development III*. WIT Press, Southampton, pp 811–820
- Goulias KG (2002) Multilevel analysis of daily time use and time allocation to activity types accounting for complex covariance structures using correlated random effects. *Transportation* 29(1):31–48
- Goulias KG (2003) *Transportation systems planning*. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 1-1 to 1-45
- Goulias KG, Kim T (2003) A longitudinal analysis of the relationship between environmentally friendly modes, weather conditions, and information-telecommunications technology market penetration. In: Tiezzi E, Brebbia CA, Uso JL (eds) *Ecosystems and sustainable development*, vol 2. WIT Press, Southampton, pp 949–958
- Goulias KG, Kim T (2005) An analysis of activity type classification and issues related to the with whom and for whom questions of an activity diary. In: Timmermans H (ed) *Progress in activity-based analysis*. Elsevier, Amsterdam, pp 309–334
- Goulias KG, Kitamura R (1992) Travel demand analysis with dynamic microsimulation. *Transp Res Rec* 1607:8–18
- Goulias KG, Kitamura R (1997) Regional travel demand forecasting with dynamic microsimulation models. In: Golob T, Kitamura R, Long L (eds) *Panels for transportation planning: methods and applications*. Kluwer, Boston, pp 321–348
- Goulias KG, Litzinger T, Nelson J, Chalamgari V (1993) A study of emission control strategies for Pennsylvania: emission reductions from mobile sources, cost effectiveness, and economic impacts. Final report to the low emissions vehicle commission. PTI 9403. The Pennsylvania Transportation Institute, University Park
- Goulias KG, Kim T, Pribyl O (2003) A longitudinal analysis of awareness and use for advanced traveler information systems. Paper to be presented at the European Commission workshop on behavioural responses to ITS – 1–3 April 2003, Eindhoven
- Goulias KG, Zekkos M, Eom J (2004) CentreSIM3 scenarios for the South Central Centre County transportation study. CentreSIM3 report submitted to McCormick Taylor Associates and the Mid-Atlantic Universities Transportation Center, April 2004, University Park

- Goulias KG, Blain L, Kilgren N, Michalowski T, Murakami E (2007) Catching the next big wave: are the observed behavioral dynamics of the baby boomers forcing us to rethink regional travel demand models? Paper presented at the 86th transportation research board annual meeting, 21–25 January 2007, Washington, DC and included in the CD ROM proceedings
- Greene WH (1997) *Econometric analysis*, 3rd edn. Prentice Hall, Upper Saddle River
- Grieving S, Kemper R (1999) Integration of transport and land use policies: state of the art. Deliverable 2b of the project TRANSLAND, 4th RTD Framework Programme of the European Commission
- Haab TC, Hicks RL (1997) Accounting for choice set endogeneity in random utility models of recreation demand. *J Environ Econ Manag* 34:127–147
- Hagerstrand T (1970) What about people in regional science? *Pap Reg Sci Assoc* 10:7–21
- Hato E (2006) Development of behavioral context addressable loggers in the shell for travel activity analysis. Paper presented at the IATBR conference, Kyoto
- Hayes-Roth B, Hayes-Roth F (1979) A cognitive model of planning. *Cogn Sci* 3:275–310
- Henson K, Goulias KG (2006) Preliminary assessment of activity and modeling for homeland security applications. *Transportation Research Record: J Transportation Research Board*, No. 1942. Transportation Research Board of the national academies, Washington, DC, pp 23–30
- Henson K, Goulias KG, Golledge R (2006) An assessment of activity-based modeling and simulation for applications in operational studies, disaster preparedness, and homeland security. Paper presented at the IATBR conference, Kyoto
- Horowitz JL (1991) Modeling the choice of choice set in discrete-choice random-utility models. *Environ Plan A* 23:1237–1246
- Horowitz JL, Louviere JJ (1995) What is the role of consideration sets in choice modeling? *Int J Res Marketing* 12:39–54
- Huigen PPP (1986) Binnen of buiten bereik?: Een sociaal-geografisch onderzoek in Zuidwest-Friesland, *Nederlandse Geografische Studies* 7. University of Utrecht, Utrecht
- Hutchinson BG (1974) *Principles of urban transport systems planning*. Scripta, Washington, DC
- JHK & Associates, Clough, Harbour & Associates, Pennsylvania Transportation Institute, Bogart Engineering (1996) Scranton/Wilkes-Barre area strategic deployment plan. Final report. Prepared for Pennsylvania Department of Transportation District 4-0. August 1996, Berlin
- Joh C-H, Arentze T, Timmermans H (2004) Activity-travel scheduling and rescheduling decision processes: empirical estimation of aurora model. *Transp Res Rec* 1898:10–18
- Jones PM, Dix MC, Clarke MI, Heggie IG (1983) *Understanding travel behaviour*. Gower, Aldershot
- Jones P, Koppelman F, Orfeuil J (1990) Activity analysis: state-of-the-art and future directions. In: Jones P (ed) *Developments in dynamic and activity-based approaches to travel analysis. A compendium of papers from the 1989 Oxford conference*. Avebury, Gower-Aldershot, pp 34–55
- Jonnalagadda N, Freedman J, Davidson WA, Hunt JD (2001) Development of microsimulation activity-based model for San Francisco. *Transp Res Rec* 1777:25–35
- Kahneman D, Tversky A (1979) Prospect theory: an analysis of decisions under risk. *Econometrica* 47(2):263–291
- Kawakami S, Isobe T (1989) Development of a travel-activity scheduling model considering time constraint and temporal transferability test of the model. In: *Transport policy, management and technology towards 2001: selected proceedings of the fifth world conference on transport research*, vol 4. Western Periodicals, Ventura, pp 221–233
- Kharoufeh JP, Goulias KG (2002) Nonparametric identification of daily activity durations using kernel density estimators. *Transp Res B Methodol* 36:59–82
- Kitamura R (1988) An evaluation of activity-based travel analysis. *Transportation* 15:9–34
- Kitamura R (1997) Applications of models of activity behavior for activity-based demand forecasting. In: Engelke LJ (ed) *Activity-based travel forecasting conference: summary, recommendations and compendium of papers*. Report of the travel model improvement program. Texas Transportation Institute, Arlington, pp 119–150
- Kitamura R (2000) Longitudinal methods. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 113–128
- Kitamura R, Fujii S (1998) Two computational process models of activity-travel choice. In: Garling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam, pp 251–279
- Kitamura R, Pas EI, Lula CV, Lawton TK, Benson PE (1996) The sequenced activity simulator (SAMS): an integrated approach to modeling transportation, land use and air quality. *Transportation* 23:267–291
- Kitamura R, Chen C, Pendyala RM (1997) Generation of synthetic daily activity-travel patterns. *Transp Res Rec* 1607:154–162
- Koppelman FS, Sethi V (2000) Closed-form discrete-choice models. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 211–225
- Krizek KJ, Johnson A (2003) Mapping of the terrain of information and communications technology (ICT) and household travel. *Transportation Research Board annual meeting CDROM*, Washington, DC
- Kuhnau JL (2001) Activity-based travel demand modeling using spatial and temporal models in the urban transportation planning system. MS Thesis, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
- Kuhnau JL, Goulias KG (2002) Centre SIM: hour-by-hour travel demand forecasting for mobile source emission estimation. In: Brebbia CA, Zannetti P (eds) *Development and application of computer techniques to environmental studies IX*. WIT Press, Southampton, pp 257–266

- Kuhnau JL, Goulias KG (2003) Centre SIM: first-generation model design, pragmatic implementation, and scenarios. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 16-1-16-14
- Kulkarni A, McNally MG (2001) A microsimulation of daily activity patterns. Paper presented at the 80th annual meeting of the Transportation Research Board, Washington, DC, 7-11 January 2001
- Kwan M-P (1994) A GIS-based model for activity scheduling in intelligent vehicle highway systems (IVHS). Unpublished PhD, Department of Geography, University of California Santa Barbara, Santa Barbara
- Kwan M-P (1997) GISICAS: an activity-based travel decision support system using a GIS-interfaced computational-process model. In: Ettema DF, Timmermans HJP (eds) *Activity based approaches to travel analysis*. Elsevier Science, New York, pp 263-282
- Lenntorp B (1976) Paths in space-time environment: a time geographic study of possibilities of individuals. *Lund Studies in Geography, Series B Human Geography*, vol 44. The Royal University of Lund, Department of Geography, Lund
- Lomborg B (2001) *The skeptical environmentalist: measuring the real state of the world*. Cambridge University Press, Cambridge
- Longford NT (1993) *Random coefficient models*. Clarendon Press, Oxford
- Los Alamos National Laboratory (2003) TRANSIMS: Transportation analysis system (Version 3.1). LA-UR-00-1725
- Loudon WR, Dagang DA (1994) Evaluating the effects of transportation control measures. In: Wholley TF (ed) *Transportation planning and air quality II*. American Society of Civil Engineers, New York
- Louviere JJ, Hensher DA, Swait JD (2000) *Stated choice methods: analysis and application*. Cambridge University Press, Cambridge
- Ma J (1997) An activity-based and micro-simulated travel forecasting system: a pragmatic synthetic scheduling approach. Unpublished PhD Dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
- Mahmassani HS, Herman R (1990) Interactive experiments for the study of tripmaker behaviour dynamics in congested commuting systems. In: *Developments in dynamic and activity-based approaches to travel analysis. A compendium of papers from the 1989 Oxford Conference*. Avebury
- Mahmassani HS, Jou R-C (1998) Bounded rationality in commuter decision dynamics: incorporating trip chaining in departure time and route switching decisions. In: Garling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam
- Manheim ML (1979) *Fundamentals of transportation systems analysis, vol 1: Basic concepts*. MIT Press, Cambridge
- Marker JT, Goulias KG (2000) Framework for the analysis of grocery teleshopping. *Transp Res Rec* 1725:1-8
- McFadden D (1998) Measuring willingness-to-pay for transportation improvements. In: Garling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam, pp 339-364
- McNally MG (2000) The activity-based approach. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 113-128
- Meyer MD, Miller EJ (2001) *Urban transportation planning*, 2nd edn. McGraw Hill, Boston
- Miller EJ (2003) Land use: transportation modeling. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 5-1 to 5-24
- Miller EJ (2006) Resource paper on integrated land use-transportation models. IATBR, Kyoto
- Miller JS, Demetsky MJ (1999) Reversing the direction of transportation planning process. *ASCE J Transp Eng* 125(3):231-237
- Miller EJ, Roorda MJ (2003) A prototype model of household activity/travel scheduling. *Transp Res Rec* 1831:114-121
- Mokhtarian PL (1990) A typology of relationships between telecommunications and transportation. *Transp Res A* 24(3):231-242
- National Cooperative Highway Research Program (2000) Report 446. Transp Res Board, Washington, DC
- Newell A, Simon HA (1972) *Human problem solving*. Prentice Hall, Englewood Cliffs
- Niemeier DA (2003) Mobile source emissions: an overview of the regulatory and modeling framework. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 13-1-13-28
- Ortuzar JD, Willumsen LG (2001) *Modelling transport*, 3rd edn. Wiley, Chichester
- Paaswell RE, Roupail N, Sutaria TC (eds) (1992) *Site impact traffic assessment. Problems and solutions*. ASCE, New York
- Patten ML, Goulias KG (2001) Test plan: motorist survey - evaluation of the Pennsylvania turnpike advanced travelers information system (ATIS) project, phase III PTI-2001-23-I. April 2001. University Park
- Patten ML, Hallinan MP, Pribyl O, Goulias KG (2003) Evaluation of the Smartraveler advanced traveler information system in the Philadelphia metropolitan area. Technical memorandum. PTI 2003-33. March 2003. University Park
- Payne JW, Bettman JR, Johnson EJ (1993) *The adaptive decision maker*. Cambridge University Press, Cambridge
- Pendyala R (2003) Time use and travel behavior in space and time. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 2-1-2-37
- Pendyala RM, Kitamura R, Kikuchi A, Yamamoto T, Fujii S (2005) The Florida activity mobility simulator (FAMOS): an overview and preliminary validation results. Presented at the 84th annual transportation research board conference and CD-ROM

- Pribyl O (2004) A microsimulation model of activity patterns and within household interactions. PhD dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
- Pribyl O (2007) Computational intelligence in transportation: short user-oriented guide. In: Goulias KG (ed) *Transport science and technology*. Elsevier, Amsterdam, pp 37–54
- Pribyl O, Goulias KG (2003) On the application of adaptive neuro-fuzzy inference system (ANFIS) to analyze travel behavior. Paper presented at the 82nd transportation research board meeting and included in the CDROM proceedings and accepted for publication in the *Transportation Research Record*, Washington DC, January 2003
- Pribyl O, Goulias KG (2005) Simulation of daily activity patterns. In: Timmermans H (ed) *Progress in activity-based analysis*. Elsevier Science, Amsterdam, pp 43–65
- Ramadurai G, Srinivasan KK (2006) *Transportation Research Board of the National Academies*, Washington, DC, pp 43–52
- Recker WW (1995) The household activity pattern problem: general formulation and solution. *Transp Res B* 29:61–77
- Recker WW, McNally MG, Root GS (1986a) A model of complex travel behavior: part I – theoretical development. *Transp Res A* 20(4):307–318
- Recker WW, McNally MG, Root GS (1986b) A model of complex travel behavior: part II – an operational model. *Transp Res A* 20(4):319–330
- Richardson A (1982) Search models and choice set generation. *Transp Res Part A* 16(5–6):403–416
- Robinson J (1982) Energy backcasting: a proposed method of policy analysis. *Energ Policy* 10(4):337–344
- Rubinstein A (1998) *Modeling bounded rationality*. MIT Press, Cambridge
- Sadek AW, El Dessouki WM, Ivan JI (2002) Deriving land use limits as a function of infrastructure capacity. Final report, project UVMR13-7, New England Region One University Transportation Center. MIT, Cambridge
- Salomon I (1986) Telecommunications and travel relationships: a review. *Transp Res A* 20(3):223–238
- Salvini P, Miller EJ (2003) ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems. Paper presented at the 10th international conference on travel behaviour research, Lucerne, August 2003
- Savage LJ (1954) *The foundations of statistics*. Reprinted version in 1972 by Dover Publications, New York
- Searle SR, Casella G, McCulloch CE (1992) *Variance components*. Wiley, New York
- Simma A, Axhausen KW (2001) Within-household allocation of travel-The case of Upper Austria. *Transportation Research Record: J Transportation Research Board*, No. 1752, TRB, National Research Council, Washington, DC 69–75
- Simon HA (1983) Alternate visions of rationality. In: Simon HA (ed) *Reason in human affairs*. Stanford University Press, Stanford, pp 3–35
- Simon HA (1997) *Administrative behavior*, 4th edn. The Free Press, New York
- Southworth F (2003) Freight transportation planning: models and methods. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 4.1–4.29
- Sparmann U (1980) Ein verhaltensorientiertes Simulationsmodell zur Verkehrsprognose. *Schriftenreihe des Instituts für Verkehrswesen* 20. Universität (TH) Karlsruhe, Karlsruhe
- Stefan KJ, McMillan JDP, Hunt JD (2005) An urban commercial vehicle movement model for Calgary. Paper presented at the 84th transportation research board meeting, Washington, DC
- Stopher PR (1994) Predicting TCM responses with urban travel demand models. In: Wholley TF (ed) *Transportation planning and air quality II*. American Society of Civil Engineers, New York
- Stopher PR, Meyburg AH (eds) (1976) *Behavioral travel-demand models*. Lexington Books, Lexington
- Stopher PR, Hartgen DT, Li Y (1996) SMART: simulation model for activities, resources and travel. *Transportation* 23:293–312
- Sundararajan A, Goulias KG (2002) Demographic microsimulation with DEMOS 2000: design, validation, and forecasting. In: Goulias KG (ed) *Transportation systems planning: methods and applications*. CRC Press, Boca Raton, pp 14-1–14-23
- Swait J, Ben-Akiva M (1987a) Incorporating random constraints in discrete models of choice set generation. *Transp Res Part B* 21(2):91–102
- Swait J, Ben-Akiva M (1987b) Empirical test of a constrained choice discrete model: mode choice in Sao Paulo, Brazil. *Transp Res Part B* 21(2):103–115
- Teodorovic D, Vukadinovic K (1998) *Traffic control and transport planning: a fuzzy sets and neural networks approach*. Kluwer, Boston
- Thill J (1992) Choice set formation for destination choice modeling. *Progr Human Geogr* 16(3):361–382
- Tiezzi E (2003) *The end of time*. WIT Press, Southampton
- Timmermans H (2003) The saga of integrated land use-transport modeling: how many more dreams before we wake up? Conference keynote paper at the Moving through net: the physical and social dimensions of travel. 10th international conference on travel behaviour research, Lucerne, 10–15, August 2003. In: *Proceedings of the meeting of the International Association for Travel Behavior Research (IATBR)*. Lucerne
- Timmermans H (2006) Analyses and models of household decision making processes. Resource paper in the CDROM proceedings of the 11th IATBR international conference on travel behaviour research, Kyoto
- Timmermans H, Arentze T, Joh C-H (2001) Modeling effects of anticipated time pressure on execution of activity programs. *Transp Res Rec* 1752:8–15
- Train KE (2003) *Discrete choice methods with simulation*. Cambridge University Press, Cambridge
- Transportation Research Board (1999) *Transportation, energy, and environment. Policies to promote sustainability*. Transportation research circular 492. TRB, Washington, DC

- Transportation Research Board (2002) Surface transportation environmental research: a long-term strategy. Transportation Research Board, Washington, DC
- Tversky A (1969) Intransitivity of preferences. *Psychol Rev* 76:31–48
- Tversky A (1972) Elimination by aspects: a theory of choice. *Psychol Rev* 79:281–299
- Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. *J Risk Uncertain* 9:195–230
- US Government (2006) Analytical perspectives. Budget of the United States Government, Fiscal year 2007. US Government printing Office, Washington, DC
- Van der Hooft T (1997) Practitioner's future needs. Paper presented at the conference on transport surveys, raising the standard. Grainau, Germany, May 24–30
- Van Middelkoop M, Borgers AWJ, Timmermans H (2004) Merlin. *Transp Res Rec* 1894:20–27
- Vause M (1997) A rule-based model of activity scheduling behavior. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Elsevier Science, New York, pp 73–88
- Veldhuisen J, Timmermans H, Kapoen L (2000) RAMBLAS: a regional planning model based on the microsimulation of daily activity travel patterns. *Transp Res A* 32:427–443
- Vovsha P, Petersen E (2005) Escorting children to school: statistical analysis and applied modeling approach. *Transp Res Rec: J Transp Res Board* 1921. Transportation Research Board of the National Academies, Washington, DC, pp 131–140
- Vovsha P, Peterson EJ, Donnelly R (2002) Microsimulation in travel demand modeling: lessons learned from the New York best practice mode. *Transp Res Rec* 1805:68–77
- Vovsha P, Peterson EJ, Donnelly R (2003) Explicit modeling of joint travel by household members: statistical evidence and applied approach. *Transp Res Rec* 1831:1–10
- Waddell P, Ulfarsson GF (2003) Dynamic simulation of real estate development and land prices within an integrated land use and transportation model system. Presented at the 82nd annual meeting of the transportation research board, 12–16 January 2003, Washington, DC. Also available in <http://www.urbansim.org/papers/>. Accessed April 2003
- Wang D, Timmermans H (2000) Conjoint-based model of activity engagement, timing, scheduling, and stop pattern formation. *Transp Res Rec* 1718:10–17
- Weiland RJ, Purser LB (2000) Intelligent transportation systems. In: *Transportation in the new millennium. State of the art and future directions. Perspectives from transportation research board standing committees*. Transportation Research Board. National Research Council. The National Academies, Washington, DC, p 6. Also in <http://nationalacademies.org/trb/>
- Wen C-H, Koppelman FS (2000) A conceptual and methodological framework for the generation of activity-travel patterns. *Transportation* 27:5–23
- Williams HCWL, Ortuzar JD (1982) Behavioral theories of dispersion and the mis-specification of travel demand models. *Transp Res B* 16(3):167–219
- Wilson EO (1998) *Consilience, the unity of knowledge*. Vintage Books, New York
- Wolf J, Guensler R, Washington S, Frank L (2001) Use of electronic travel diaries and vehicle instrumentation packages in the year 2000. Atlanta regional household travel survey. Transportation research circular, E-C026, March 2001. Transportation Research Board, Washington, DC
- Zhang J, Timmermans HJP, Borgers AWJ (2005) A model of household task allocation and time use. *Transp Res B* 39:81–95

Modelling of Pedestrian and Evacuation Dynamics

Mohcine Chraïbi¹, Antoine Tordeux²,
Andreas Schadschneider³ and Armin Seyfried^{1,4}

¹Institute for Advanced Simulation,
Forschungszentrum Jülich GmbH, Jülich,
Germany

²School of Mechanical Engineering and Safety
Engineering, University of Wuppertal,
Wuppertal, Germany

³Institut für Theoretische Physik, Universität zu
Köln, Köln, Germany

⁴School of Architecture and Civil Engineering,
University of Wuppertal, Wuppertal, Germany

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Classification of Models

Acceleration-Based Models

Velocity-Based Models

Velocity Obstacle Models

Other Velocity-Based Models

Rule-Based Models

Macroscopic Models

Future Directions and Open Problems

Bibliography

Glossary

Pedestrian A pedestrian is a person travelling on foot. In this article, other characterizations are used, depending on the context, e.g., agent or particle.

Crowd A large group of pedestrians who have gathered together. Depending on the perspective, more specific definitions exist.

Microscopic models Microscopic models represent each pedestrian separately with individual properties (e.g., walking velocity or route choice

behavior) and his/her interactions with other individuals. Typical models that belong to this class are cellular automata and the force-based models.

Macroscopic models Macroscopic models do not distinguish individuals. The description is based on aggregate quantities, e.g., appropriate densities. Typical models belonging to this class are fluid-dynamic approaches.

Acceleration-based models Acceleration-based models are microscopic models defined by a system of *second-order* ordinary differential equations. The resulting acceleration is integrated twice to calculate the velocity and position of modelled pedestrians.

Velocity-based models Velocity-based models are microscopic models defined by means of a system of *first-order* ordinary differential equations. The resulting velocity is integrated once to calculate position of modelled pedestrians.

Rule-based (decision-based) models Rule-based or decision-based models are not based on differential equations. Instead the dynamics is based on rules or decisions that the agents make in order to determine their new positions, velocities, etc. Typically time is considered to be discrete in this approach, i.e., it increases in certain discrete time steps Δt at which decisions are made.

Cellular automata model Cellular automata are models in which space and time are discrete and state variables take on finite set of discrete values. Usually, a regular square lattice with cells of size 40 cm^2 is used. However, hexagonal lattices have been used too. A discrete time is usually realized through the *parallel* or *synchronous update* where all particles or sites are moved at the same time. This introduces a timescale.

Fundamental diagram In traffic engineering (and physics): Density dependence of the flow, i.e., a function $J(\rho)$. Due to the hydrodynamic relation $J = \rho v b$, equivalent representations used frequently are $v = v(\rho)$ or $v = v(J)$. The fundamental diagram is probably the most important quantitative characterization of traffic systems.

Lane formation In bidirectional flows often dynamically lanes are formed in which all pedestrians move in the same direction.

Bottleneck A bottleneck is in general a limited resource for pedestrian flows. This can be, for example, a door, a narrowing in a corridor, or stairs, i.e., locations of reduced capacity. At bottlenecks jamming occurs if the inflow is larger than the capacity.

Definition of the Subject and Its Importance

Safety in public spaces, especially at mass events where thousands of people gather in a restricted area, is a major challenge for organizers, security staff, and other stakeholders. Time and again, overcrowding, heavy congestions, and pushing in crowds lead to injuries or even fatalities. To name only few recent incidents, we refer to the Loveparade in Duisburg, Germany, the Hajji in Mecca, or several large religious gatherings in India. Therefore, in recent years we noticed an increasing activity in research on pedestrian and evacuation dynamics. Besides an improved qualitative and quantitative understanding of the general phenomena observed in large human crowds, especially the development of accurate models for pedestrian streams, has become a very active field of research. After validation and calibrations, these models can be used for planning of large events, the design of transport infrastructures, or the optimization of evacuation routes. These important applications have led to the development of many different models and simulation tools.

In this contribution we will discuss general aspects of the various modelling approaches that have been used for the description of pedestrian streams. We use a classification that allows to distinguish various model classes and discuss the most important representatives of each class in more detail.

Introduction

Models in pedestrian dynamics describe the movement of single pedestrians, small groups, or the assembling and dispersal of large gatherings

which could happen, e.g., outdoor in a city or inside of buildings. They are developed in natural science, computer science, engineering, social sciences, or psychology. The scope and the type of the models are closely related to the scientific area or the application and thus are diverse.

In physics, pedestrians are treated as driven particles to study the emergence of collective phenomena like stop-and-go waves, lane formation, or clogging and intermittent flow at bottlenecks. The objective of minimal modelling approaches is to identify basic microscopic mechanisms and types of interaction that are relevant for the emergence of the observed collective phenomena. In traffic engineering, models focus on transport characteristics like flow, density, speed, or route choices. In fire safety science, the process of building evacuation including behavioral factors like the reaction time or the influence of smoke and fire is addressed. The models support planners in the design and dimensioning of pedestrian facilities in dense cities, high occupancy buildings, infrastructures for public transport, or large-scale events outside. For an overview we refer to the conference series “Traffic and Granular Flow” and “Pedestrian and Evacuation Dynamics” (see section “Books and Reviews”). In computer animation, the aim of the models is a realistic appearance of moving agents, while in robotics the agents should steer in complex environment safely without collisions with other agents or moving objects; see, e.g., Ali et al. (2013). In psychology, however, it is of interest how pedestrians perceive information in complex environments and how they gain spatial knowledge using it for navigation and wayfinding (Chrastil and Warren 2015; Rio et al. 2014).

The perspective on the real system illustrates that moving pedestrians constitute a system with many interesting relations as well as complex phenomena which could depend on a multitude of environmental, structural, and human factors. Looking on the level of abstraction of actual models, one has to note that pedestrians are three-dimensional objects and a complete description of their highly developed and complicated motion sequence is rather difficult. Most models in pedestrian dynamics treat them as points or two-dimensional objects considering the vertical

projection of the body. Besides the complicated locomotor systems, pedestrians are equipped with many senses, cognitive abilities, and memory. Considering the modelling of the space of movement, various factors must be taken into account, including the type of building and its facilities, like rooms, corridors, stairs, or ramps with different spatial structures, lighting, temperatures, and so on. Pedestrians, as humans, are also subject of psychology and sociology. Many concepts and terms used by engineers and natural scientists to model pedestrian dynamics are borrowed from these research fields. Often this is connected with an oversimplification; see the discussion in Sieben et al. (2017) and Templeton et al. (2015). Psychology considers single pedestrians and their behavior and experiences. In pedestrian dynamics, the influence of senses like viewing, hearing, or the tactile sense on the movement and cognitive abilities like wayfinding are of interest. In sociology, pedestrians are treated as entities, belonging to a group (families or friends) sharing a social identity (e.g., protester or police) or following social norms (e.g., waiting in a queue instead of pushing toward an overloaded exit). Thus many norms and rules should be considered as well for a realistic modelling of the movement of crowds, groups, or single pedestrians.

The high complexity of the system in combination with the very different scopes of modelling illustrates why in pedestrian dynamics many different approaches for models are developed and why a huge number of model variants exist. Often

a model or a variant of a model is introduced to describe a special phenomenon or to consider an application-driven influence on the movement. It also explains the lacking generality of actual approaches.

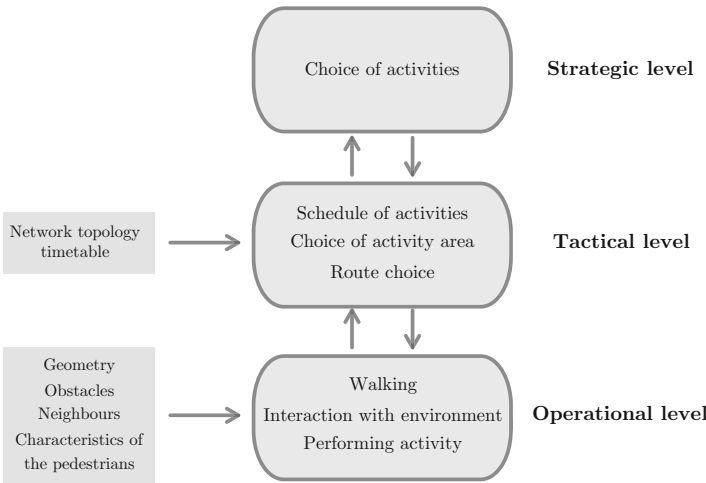
Classification of Models

An often used categorization in pedestrian dynamics takes three different levels of behavior into account (Fig. 1). At the *strategic level*, pedestrians decide which activities they like to perform and the order of these activities. With the choices made at the strategic level, the *tactical level* concerns the short-term decisions made by the pedestrians, e.g., choosing the route taking into account obstacles, density of pedestrians, etc. Finally, the *operational level* describes the actual walking behavior of pedestrians, e.g., their immediate decisions necessary to avoid collisions, etc.

The processes at the strategic and tactical level are usually considered to be exogenous to the pedestrian simulation. Here information from other disciplines (sociology, psychology, etc.) is required. In the following we will mostly be concerned with the operational level, although some of the models that we are going to describe allow to take into account certain elements of the behavior at the tactical level as well.

Modelling on the operational level is usually based on variations of models from physics.

Modelling of Pedestrian and Evacuation Dynamics, Fig. 1 The different levels of modelling of pedestrian behavior. (According to Daamen (2004) and Hoogendoorn et al. (2002))



Indeed the motion of pedestrian crowds has certain similarities with fluids or the flow of granular materials. The goal is to find models which are as simple as possible but at the same time can reproduce “realistic” behavior in the sense that the empirical observations are reproduced. Therefore, based on the experience from physics, pedestrians are often modelled as simple “particles” that interact with each other. Extensions to models of “intelligent” particles with numerous parameters and mechanisms for the interaction with the neighbors and the environment are usually called *multi-agent systems* see, e.g., Kienar et al. (2016a) and Ondrej and Pettré (2010)). Agents interact with each other and with the environment. They can have an internal state reflecting, e.g., their goals and general behavior.

There are several characteristics which can be used to classify the modelling approaches:

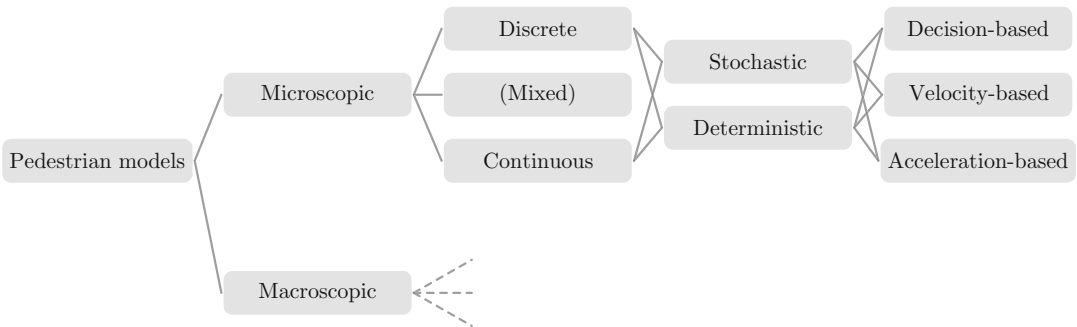
- **Microscopic vs. macroscopic:** In microscopic models each individual is represented separately which allows to introduce different types of pedestrians with individual properties in a natural way. In macroscopic models different individuals cannot be distinguished. Instead the state of the system is described by time and space averages of density, velocity, and flow.
- **Discrete vs. continuous:** Each of the three basic variables for a description of a system of pedestrians, namely, space, time, and state variable (e.g., velocities), can be either discrete (i.e., an integer number) or continuous (i.e., a real number). Here all combinations are possible. In a cellular automaton approach, all variables are by definition discrete, whereas in hydrodynamic models all are continuous. These are the most common choices, but other combinations are used as well. A discrete time is usually realized through the *parallel* or *synchronous update* where all particles are moved at the same time. This introduces a timescale.
- **Deterministic vs. stochastic:** The dynamics of pedestrians can either be modelled deterministically or stochastically. In the first case, the movement at a certain time is completely determined by the present state. In stochastic models, the movement is determined by probabilities such that the agents can react differently in the same situation. Such stochastic behavioral rules often generate a rather realistic representation of complex systems like pedestrian crowds. On the other hand, the stochasticity in the models reflects the lack of knowledge of the underlying physical processes that, e.g., determine the decision-making of the pedestrians.
- **Rule-based vs. acceleration-based vs. velocity-based:** The motion of the pedestrians to a new position can be determined in different ways. In a large class of models, first the new velocity or acceleration is determined which then allows (by integration) to calculate the new position of a pedestrian. Usually this requires the solution of a first-order and second-order differential equation. Such models can be classified as velocity-based and acceleration-based models, respectively. In most models that do not belong to these two classes, the determination of the new position does not require the solution of a differential equation. Instead the new position is specified by certain rules which, e.g., take into account the desired direction of motion and the position of other pedestrians and obstacles. This can be seen most clearly in cellular automata models. Often in these models, time is taken as a discrete variable. Models in this third class can be called rule-based models or decision-based models. They focus on the *intrinsic properties* of the agents, and thus the rules are often justified from psychology. It should be mentioned that sometimes a clear distinction between velocity-

and acceleration-based models on one hand and rule-based models on the other cannot always be made and several models combine aspects of both approaches.

In acceleration-based models, the basic interactions are typically described by forces: agents “feel” a force exerted by others and the infrastructure. It is a physical approach based on the observation that the presence of others leads to deviations from a straight motion. In analogy to Newtonian mechanics, this change of speed, i.e., the acceleration, is explained by a force. The equations of motion allow to determine the velocity and new position of each agent by integration or solving second-order differential equations. In velocity-based models, the speed and the direction of an agent depend on the distance and position of the surrounding neighbors and obstacles. The equations of motion determine directly the velocity without referring to an acceleration. Therefore they are typically given by first-order differential equations. They are generally based on collision avoidance and speed optimization principles. In the simplest models, the speed is directly proportional to the distance to the next agent or obstacle in front, while the direction is the one maximizing the speed to the desired destination. In contrast to rule-based models, acceleration- and velocity-based models emphasize the *extrinsic properties* and their relevance for the motion of the agents.

- **Heuristic vs. first-principles:** Heuristic models typically include several interaction terms considered to be relevant. These interactions are defined by parameters of the model that are used to fit empirical data. First-principle models are derived from certain postulates considered to be fundamental. Often a clear distinction between the two approaches is not possible.
- **High vs. low fidelity:** *Fidelity* here refers to the apparent realism of the modelling approach. High-fidelity models try to capture the complexity of decision-making, actions, etc. that constitute pedestrian motion in a realistic way. In contrast, in the simplest models, pedestrians are represented by particles without any stimulus-response mechanism like viewing or hearing and cognitive capabilities like spatial knowledge or decision-making. The introduction of “internal states” of an agent which influence their behavior can be interpreted as some kind of “intelligence” that leads to more complex approaches, like multi-agent models.

Fig. 2 shows the variety of model classes obtained from this classification. There are examples for almost every class shown here. It should be mentioned that a clear classification according to the characteristics outlined here is not always possible. In the following we will describe some model classes in more detail.



Modelling of Pedestrian and Evacuation Dynamics, Fig. 2 Different possible combinations of characteristics of the microscopic pedestrian models. Based on the proposed model classification, a large number of different approaches can be distinguished. Here we show only

(part of) the classification scheme for microscopic models, but macroscopic models present similar classification. The most common approaches are microscopic discrete, stochastic, rule-based models and continuous, deterministic acceleration-based models

Acceleration-Based Models

A popular class of microscopic models for pedestrian dynamics is defined by the so-called acceleration-based, also known as force-based, models, where the movement of pedestrians is defined by a superposition of exterior forces. Given initial values of the state variables, the resulting second-order ordinary differential equations can be solved numerically to obtain the trajectories of pedestrians.

In general, the acceleration of pedestrian i at a given time step is directly proportional to a driving force, pointing to a specific target. In most known acceleration-based models, this force model is an exponential acceleration from zero to a certain desired speed. However, in the presence of other pedestrians or static obstacles, the trajectory of pedestrian i deviates from a straight line, due to different repulsive forces acting on i from its environment.

For the solution of the equations of motion of Newtonian many-particle systems, the well-founded molecular dynamics technique exists. However, in most studies so far, the distinctions between pedestrian and Newtonian dynamics with respect to the principles *actio equals reactio*, superposition of forces, or the meaning of the mass in the second law are not discussed in detail.

The mathematical expression of the repulsive force differs from one model to another. In general, its magnitude increases with decreasing inter-distance of two pedestrians. This assures a certain volume exclusion of pedestrians, although in contrast to cellular automata models 6.1, pedestrians may still collide or even overlap with each other. In section “[Limitations of Acceleration-Based Models](#)” numerical difficulties and properties of acceleration-based models will be discussed in more details.

In the following we give a brief overview of some important milestones in the development of acceleration-based models.

Heraï-Tarui Model (HTM)

In 1975 Hirai and Tarui proposed the first known microscopic force-based model to investigate the movement of pedestrians in a 2D space (Hirai and

Tarui 1975). In their seminal work, the authors investigated several aspects of human’s behavioral motion. On one side, the proposed model considered the movement of pedestrians on the “operative level,” where pedestrians move in direction of an assigned goal while avoiding collisions with other pedestrians or obstacles. On the other side, the model incorporates different aspects of the “tactical level” of human behavior as well, e.g., group behavior and the influence of guiding signs on agent’s way-finding – aspects which recently caught the attention of the community of pedestrian dynamics with several emerging studies.

The equation of motion of the HTM is

$$m_i \frac{d^2 \mathbf{x}_i}{dt^2} + v_i \frac{d\mathbf{x}_i}{dt} = \mathbf{f}_{1i} + \mathbf{f}_{2i} + \boldsymbol{\xi}, \quad (1)$$

where m_i , v_i and $\mathbf{x}_i$ are the mass, coefficient of viscosity, and the position of the individual i , respectively. $\mathbf{f}_{1i}$ and $\mathbf{f}_{2i}$ are external forces acting upon the individual i , implementing different modelling concepts. $\mathbf{f}_{1i}$ is a force required by the individual i to form a group together with other individuals and to move forward, while $\mathbf{f}_{2i}$ is a force exerted by the environment around the individual i .

The mean forces $\mathbf{f}_{1i}$ and $\mathbf{f}_{2i}$ are defined as a superposition of other forces, expressing different ideas. The “group” force, for instance, consists of three components:

$$\mathbf{f}_{1i} = \mathbf{f}_{ai} + \mathbf{f}_{bi} + \mathbf{f}_{ci}, \quad (2)$$

where $\mathbf{f}_{ai}$ is a driving force causing i to move forward with constant speed. $\mathbf{f}_{bi}$ is an interaction force acting between the individual i and other individuals. Depending on the distance between two pedestrians, it can be attractive or repulsive. Finally, $\mathbf{f}_{ci}$ is the group force, which acts only if some individuals move together in a restricted domain.

Both forces $\mathbf{f}_{ai}$ and $\mathbf{f}_{ci}$ are distance and angle dependent. This expresses the observation that the influence of other pedestrians on i is greater in its direction of motion than on the side. No forces act on individuals from behind.

In analogy to the “group force,” the “environmental” force $\mathbf{f}_{2i}$ is defined as the sum of other forces, each with a specific effect, e.g., repulsion from walls and attraction toward signs. For more details the reader is referred to the original paper.

The last main term in the model ξ is a random force acting upon i to consider uncertainties in the modelled behavior. Its direction varies stochastically, and its magnitude is a function of the distance between i and the wall.

Hirai and Tarui showed that their model exhibits, to some extent, realistic evacuation behavior in a simplified train station. The authors validated their model with respect to experiments performed on rats, making the assumption that human’s behavior in “panic”-like situations may be comparable to animal’s dynamics.

Although the model was not elaborated further whether in the original paper or in the literature, it shows some original concepts that were investigated lately in different works, e.g., group behavior, route choice modelling, attraction to a site, memory effects of objects, etc.

Social Force Model and Its Derivatives

The social force model (Helbing and Molnár 1995) (SFM) is the most investigated acceleration-based

model for pedestrian dynamics, probably due to the simplicity of its definition, compared to the HTM, since it only implements the movement of pedestrians on the operative level. Besides, unlike the HTM, an exponential acceleration of i from zero to a predefined maximal speed is chosen (Pipes 1953). The main equation of motion is given by

$$\frac{d^2 \mathbf{x}_i}{dt^2} = \mathbf{f}(\mathbf{v}_i, \Delta \mathbf{v}_{ij}, \Delta x_{ij}) + \frac{\mathbf{v}_i^{(0)} - \mathbf{v}_i}{\tau}, \quad (3)$$

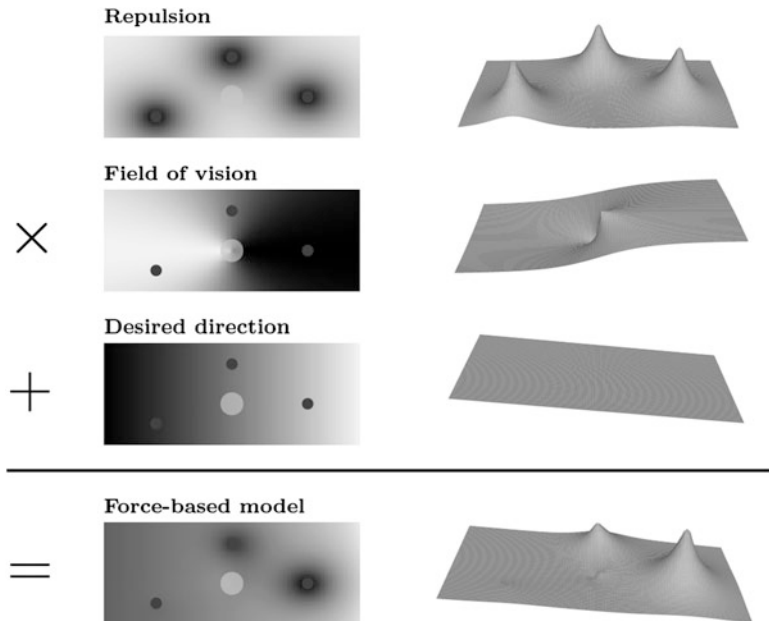
where $\mathbf{f}$ is the sum of the interactions of i with its environment, τ is a time constant, and $\mathbf{v}_i = d\mathbf{x}_i/dt$. In the original paper, these forces are described by the negative gradient of a potential field with elliptical equipotential lines.

Formally, the social force model is the sum of repulsion with the neighbours and obstacles weighted by the field of vision and a driving force for the desired direction (see Fig. 3). The mathematical expression of the driving force in Eq. (3) is systematically used in all known force-based models, as it corresponds well to human’s acceleration in free movement (Moussaïd et al. 2009).

The SFM in its original form was shown to reproduce two qualitative phenomena, namely, oscillations at bottlenecks and lane formations.

Modelling of Pedestrian and Evacuation Dynamics,

Fig. 3 Illustration for the social force model. The acceleration rate of a pedestrian is the sum of exponential repulsion weighted by the field of vision and a driving force



In high-density situations, the abovementioned “social” forces, initially defined to model navigation of pedestrians around obstacles (and other pedestrians), show that they may not be sufficient to guarantee a certain volume exclusion of pedestrians. Therefore, additional physical (contact) forces, e.g., friction and compression, were introduced later on in Helbing et al. (2002). Besides this addition, the qualitative form of the repulsive forces in the generalized SFM was assumed to be exponential. In Chen et al. (2017) a comprehensive review of the evolution of the SFM over the years is given. Therein, the authors show the different modifications and enhancements of the generic model.

Limitations of Acceleration-Based Models

The appeal of the force-based models is given mainly by the analogy to Newtonian dynamics. In particular scenarios, where a pedestrian in a crowd has limited freedom, this analogy can be justified. However, unlike particles or solid objects, pedestrians can show a complex and non-smooth dynamics while moving in space, especially in situations where prompt decisions are made, like, for example, stopping. Acceleration-based models describe particle with inertia which do not stop instantly.

Describing particles with inertia often leads to overlapping and oscillations of the modelled pedestrians. Depending on the strength of the repulsive forces, the modelled body of pedestrians can be excessively overlapped and hence violate the principle of volume exclusion. In Kretz (2015), it was shown that in an 1D-scenario, oscillations in the movement of pedestrians are *inherent*. Pedestrians do not stop and keep moving independently of the actual situation. In some situations pedestrians perform repetitive backward and forward movement due to, e.g., high repulsive forces. In Chraïbi et al. (2010) it was found that parameter of the repulsive forces can be adequately adjusted to mitigate overlapping and oscillations, without annihilating them totally. Furthermore, velocity-based models can be described by construction volume exclusion (see the next section).

In Köster et al. (2013), Köster et al. analyzed the properties of the SFM from a numerical perspective. They showed that in order to overcome the numerical difficulties that result in a straightforward implementation of the SFM, several changes in the model are necessary. Hence, the so-called mollified SFM, which removes discontinuities from the original equation of motion, shows better numerical behavior. Furthermore, it allows the usage of high-order fast-converging numerical solvers (usually an Euler scheme with a very small time step is used to numerically solve the equation of motion, which results in very high computational overhead), which leads to a decrease in simulation time and increase in the accuracy of the obtained results. With the presented model, collisions and oscillations could be mitigated as well, especially in critical situations, e.g., near intermediate goals.

Another issue of acceleration-based models, which is the superposition principle, was discussed in Seitz et al. (2016). The authors suggest a change in the modelling perspective towards other paradigms, which may inspire new models.

Velocity-Based Models

In contrast to acceleration-based models, the velocity in velocity-based (VB) models is instantaneously adjusted to the neighborhood and the environment with no inertia or an implicit implementation of a reaction time. Such a modelling approach is largely inspired by motion planning in robotics, since robots move with almost no inertia and react instantaneously to a command. Constraints on the velocity in case of contact allow to model hard-core body exclusion (volume exclusion) and to describe collision-free pedestrian dynamics with no overlapping (see Maury et al. (2008) for a general framework).

Generally speaking VB approaches are based on visual considerations (local interaction) and optimization techniques under collision avoidance constraints. Mathematically, VB models are differential equations of the first order. A typical example is

$$\frac{d\mathbf{x}_i}{dt} = \mathbf{V}((\mathbf{x}_j - \mathbf{x}_i, \mathbf{v}_j), j \in N_i). \quad (4)$$

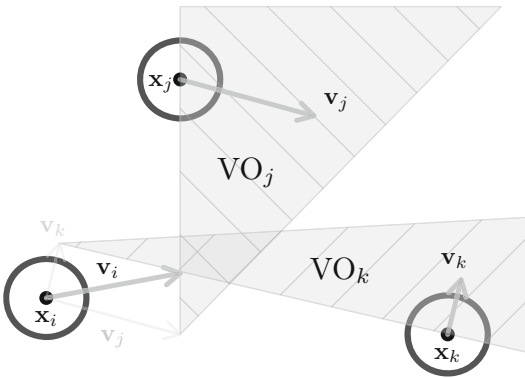
Here the velocity is a function $\mathbf{V}$ depending on the relative positions $\mathbf{x}_j - \mathbf{x}_i$ and velocities $\mathbf{v}_j$ of the neighbors $j \in N_i$ of the considered pedestrian i .

Velocity Obstacle Models

Pioneer VB models inspired from the field of robotics are *velocity obstacle models* (Fiorini and Shiller 1998). The velocity in this model class is determined by minimizing the deviation to the desired speed while avoiding collision. The collision possibilities are evaluated over horizon times by assuming constant velocities of the neighbors and carrying out the velocity obstacle cones $\text{VO}(i)$ (see Fig. 4). No collision occurs when the velocity is set outside the cones. The velocity is then the solution of the optimization problem

$$\mathbf{V}_i = \arg \min_{\mathbf{v} \notin \{\text{vo}(i)\}} \|\mathbf{v} - \mathbf{v}_0\|, \quad (5)$$

which minimizes the deviation to the desired speed $\mathbf{v}_0$. Similar optimization problems coupled to exclusion constraints in case of contact are solved in Maury and Venel (2011) for evacuation dynamics.



Modelling of Pedestrian and Evacuation Dynamics, Fig. 4 Velocity obstacle cones $\text{VO}(i)$ for the pedestrian i . No collision occurs when the velocity is set outside the cones.

Velocity obstacle models can show so-called ping-pong undesired oscillatory effects (i.e., two pedestrians switching from one direction to another). This phenomenon results from the implicit definition of velocities in the model (the speed of a pedestrian depends on the speed of the neighbor and vice versa). The *reciprocal velocity obstacle model* (RVO) allows to avoid this undesired behaviour by taking into account the fact that the neighbors make similar collision avoidance reasoning (van den Berg et al. 2008). However, this formulation only guarantees hard-core body exclusion under specific conditions. The *optimal reciprocal collision avoidance* (ORCA) overcomes this limitation and provides sufficient and efficient conditions for moving agents to avoid collisions among one another (van den Berg et al. 2011). ORCA modelling approach and its variants are frequently used to describe pedestrian dynamics (see, e.g., Guy et al. (2009, 2010a, b), Kim et al. (2014), Mordvintsev et al. (2014), and Paris et al. (2007)).

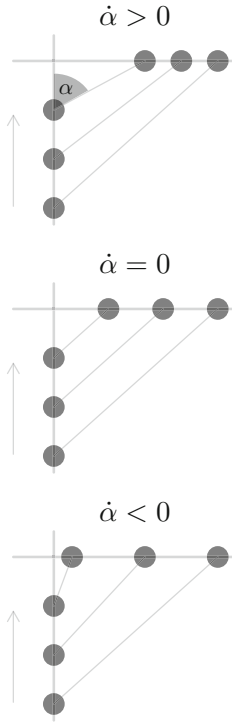
Other Velocity-Based Models

Other VB models consider the velocity as the product of a speed V (scalar) by a direction $\mathbf{e}$ (such that $\|\mathbf{e}\| = 1$):

$$\begin{aligned} \frac{d\mathbf{x}_i}{dt} &= V((\mathbf{x}_j - \mathbf{x}_i, \mathbf{v}_j), j \in N_i) \\ &\times \mathbf{e}((\mathbf{x}_j - \mathbf{x}_i, \mathbf{v}_j), j \in N_i). \end{aligned} \quad (6)$$

Both speed V and direction $\mathbf{e}$ depend on the relative positions of the neighbors and their velocities. For the *synthetic-vision-based steering model* (Ondrej and Pettré 2010), the speed V is a function of the time-to-interaction (TTI, a quantity close to the time-to-collision) with the neighbors, while the direction $\mathbf{e}$ is a function of the TTI, the bearing angle and its derivation

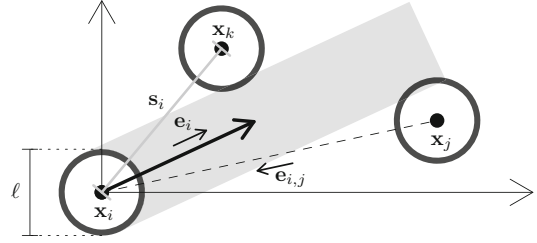
$$\begin{aligned} V_i &= v_0 \left(1 - \exp \left(- \min_j \text{TTI}_{i,j} \right) \right) \text{ and } \mathbf{e}_i \\ &= F \left(\text{TTI}_{i,j}, \alpha_{i,j}, \dot{\alpha}_{i,j} \right). \end{aligned} \quad (7)$$



Modelling of Pedestrian and Evacuation Dynamics, Fig. 5 Bearing angle α of two pedestrians. The pedestrians collide when the bearing angle is constant (i.e., when its derivative vanishes)

The bearing angle α is the angle between the current direction and the direction to the neighbour (see Fig. 5). It remains constant in case of collision, while it varies sharply for safe crossings.

In other VB approaches, the speed depends, in analogy to optimal velocity (OV) models in traffic, on the minimal distance in front (Tordeux et al. 2016), while the direction is, as in acceleration-based models, a sum of exponential repulsion with the neighbors (Dietrich and Köster 2014). If $\mathbf{e}_i$ is the direction of the pedestrian i , $\mathbf{e}_{i,j}$ the direction from j to i , and $s_{i,j} = \|\mathbf{x}_j - \mathbf{x}_i\|$ the spacing distance between pedestrians i and j , and if the pedestrians are considered as discs of diameter ℓ , then the pedestrians in front of the pedestrian i are the set $J = \{j, \mathbf{e}_i \cdot \mathbf{e}_{i,j} \leq 0 \text{ and } \mathbf{e}_i^\perp \cdot \mathbf{e}_{i,j} \leq \ell/s_{i,j}\}$. The minimum spacing in front is $s_i = \min_{j \in J} s_{i,j}$ (see Fig. 6). The optimal velocity VB model with exponential repulsion is then.



Modelling of Pedestrian and Evacuation Dynamics, Fig. 6 Minimal spacing in front. The pedestrians in front of the pedestrian i (i.e., into the gray area) are the set $J = \{j, \mathbf{e}_i \cdot \mathbf{e}_{i,j} \leq 0 \text{ and } \mathbf{e}_i^\perp \cdot \mathbf{e}_{i,j} \leq \ell/s_{i,j}\}$, while the minimum spacing in front is $s_i = \min_{j \in J} s_{i,j}$

$$V_i = f(s_i) \text{ and } \mathbf{e}_i$$

$$= \frac{1}{N} \left(\frac{\mathbf{v}_0}{\|\mathbf{v}_0\|} + A \sum_j \exp(-s_{i,j}/B) \right). \quad (8)$$

Here N is a normalization constant and f is the optimal velocity function. Any positive OV function f depending on the minimal spacing in front s_i ensures collision-free dynamics as soon as $f(s) = 0$ for all $s \leq \ell$, independently to the direction model (Tordeux et al. 2016). Note that in contrast to VB models described above, the velocity in this model solely depends on the relative positions of the neighbors, and not on their velocities, making the system of velocities explicitly defined, in contrast to the velocity obstacle or the synthetic-vision-based steering models.

Models with Noise

VB models do not incorporate relaxation or delay mechanisms and consequently do not describe stop-and-go wave phenomena. However, relaxation and diffusion processes can be introduced by using specific noises in the dynamics. For instance, in Tordeux and Schadschneider (2016), the speed is proportional to the distance spacing and has an additive noise

$$\begin{cases} \frac{dx_i}{dt} &= \frac{1}{T}(x_{i+1} - x_i - \ell) + \varepsilon_i, \\ \frac{d\varepsilon_i}{dt} &= -\frac{1}{\beta}\varepsilon_i + \alpha \frac{dW_i}{dt} \end{cases} \quad (9)$$

$i + 1$ being the pedestrian directly in front of the pedestrian i . The parameter T is the time gap and ℓ

is the pedestrian size, α and β are parameters related to the noise, and W_j is a Wiener process.

Classically the noise is white in acceleration-based models (see, e.g., (Helbing and Molnár (1995))). Here the noise is a Brownian one relaxed at the second order to describe the autocorrelation of the speed residuals. Simulation results show that the relaxed noises allow to describe oscillating traffic waves (stop-and-go dynamics) (Tordeux and Schadschneider 2016). Thus this modelling Ansatz of stop-and-go waves for pedestrian dynamics is noise-induced. It contrasts with the modelling of stop-and-go for traffic flow, generally done by means of instability and phase transitions.

Rule-Based Models

In contrast to the models discussed in sections “Acceleration-Based Models” and “Velocity-Based Models,” another class of models is not based on differential equations. Instead the dynamics is based on rules or decisions that the agents make in order to determine their new positions, velocity, etc. Therefore these models can be called *decision-based models* or *rule-based models*. Typically time is considered to be discrete in this approach, i.e., it increases in certain time steps Δt such that decisions are only made at discrete times. Since this approach does not lead to a differential equation-based description of the dynamics, these models could be interpreted as zeroth-order models. The time step corresponds to a natural timescale Δt which could, e.g., be identified with some reaction time. This can be used

for the calibration of the model which is essential for making quantitative predictions.

Cellular Automata

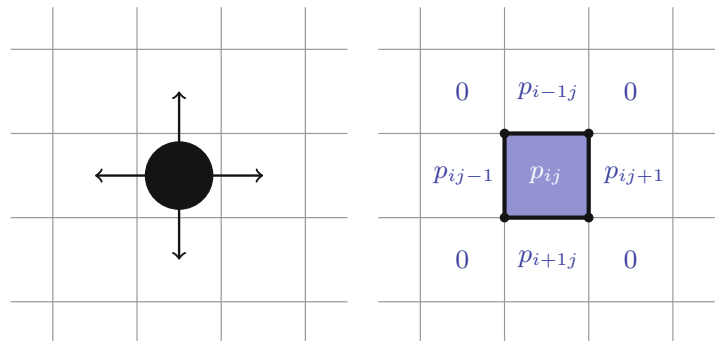
Cellular automata (CA) are rule-based dynamic models that are discrete in space, time, and state variable which in the case of traffic usually corresponds to the velocity. In computer simulations the discrete time is usually realized through a *parallel* or *synchronous* update where all pedestrians move at the same time. A natural space discretization into cells can be derived from the maximal densities observed in dense crowds which gives the minimal space requirement of one person. Usually each cell in the CA can only be occupied by one particle (exclusion principle) so that this space requirement can be identified with the cell size. In this way, a maximal density of 6.25 P/m^2 (Weidmann 1993) leads to a cell size of $40 \times 40 \text{ cm}^2$. The exclusion principle and the modelling of humans as non-compressible particles mimics short-range repulsive interactions, i.e., the “private-sphere.”

The dynamics is usually defined by rules which specify the transition probabilities for the motion to one of the neighboring cells (Fig. 7). The models differ in the specification of these probabilities as well in that of the “neighborhood.” For deterministic models all except of one probability are zero.

The first cellular automata (CA) models (Blue and Adler 2000; Fukui and Ishibashi 1999b; Klüpfel et al. 2000; Muramatsu et al. 1999) for pedestrian dynamics can be considered two-dimensional variants of the asymmetric simple exclusion process (ASEP) (for reviews, see Blythe

Modelling of Pedestrian and Evacuation

Dynamics, Fig. 7 A particle, its possible directions of motion, and the corresponding transition probabilities p_{ij} for the case of a von Neumann neighborhood



and Evans (2007), Derrida (1998), and Schütz (2001)) or models for city or highway traffic (Biham et al. 1992; Chowdhury et al. 2000; Nagel and Schreckenberg 1992) based on it. Most of these models represent pedestrians by particles without any internal degrees of freedom. They can move to one of the neighboring cells based on certain transition probabilities which are determined by three factors: (1) the desired direction of motion, e.g., to find the shortest connection, (2) interactions with other pedestrians, and (3) interactions with the infrastructure (walls, doors, etc.).

Blue-Adler Model

The model of Blue and Adler (2000, 2002) is based on a variant of the Nagel-Schreckenberg model (Nagel and Schreckenberg 1992) of highway traffic. Pedestrian motion is considered in analogy to a multilane highway. The structure of the rules is similar to the basic two-lane rules suggested in Rickert et al. (1996). The update is performed in four steps which are applied to all pedestrians in parallel. In the first step, each pedestrian chooses a preferred lane. In the second step, the lane changes are performed. In the third step, the velocities are determined based on the available gap in the new lanes. Finally, in the fourth step, the pedestrians move forward according to the velocities determined in the previous step.

In counterflow situations, head-on conflicts occur. These are resolved stochastically, and with some probability opposing pedestrians are allowed to exchange positions within one time step. Note that the motion of a single pedestrian (not interacting with others) is deterministic otherwise.

In the Blue-Adler model, motion is not restricted to nearest neighbor sites. Instead pedestrians can have different velocities $v_{\max}$ which correspond to the maximal number of cells they are allowed to move forward. In contrast to vehicular traffic, acceleration to $v_{\max}$ can be assumed to be instantaneous in pedestrian motion.

Floor Field CA

In the floor field CA (Burstedde et al. 2001, 2002; Kirchner and Schadschneider 2002; Schadschneider 2002) the transition probabilities to neighboring cells are no longer constant but

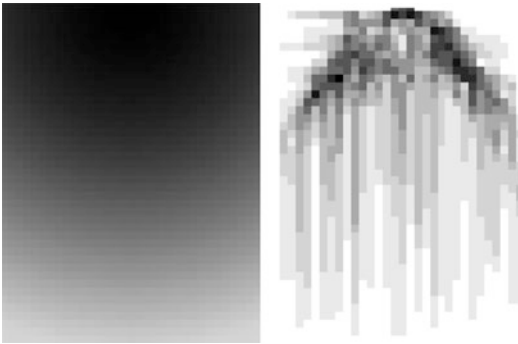
vary dynamically. This is motivated by the process of chemotaxis (see (Ben-Jacob (1997) for a review) used by some insects (e.g., ants) for communication. They create a chemical trace to guide other individuals to food sources. In the approach of Burststedde et al. (2001), the pedestrians also create a trace which, however, is only virtual although one could assume that it corresponds to some abstract representation of the path in the mind of the pedestrians. This reduces interactions to local ones that allow efficient simulations in arbitrary geometries since, e.g., it does not require to check whether the interaction between particles are screened by walls, etc. The number of interaction terms always grows linearly with the number of particles.

The translation into local interactions is achieved by the introduction of so-called *floor fields*. The transition probabilities for all pedestrians depend on the strength of the floor fields in their neighborhood in such a way that transitions in the direction of larger fields are preferred. The *dynamic floor field* D_{ij} corresponds to a virtual trace which is created by the motion of the pedestrians and in turn influences the motion of other individuals. Furthermore it has its own dynamics. Diffusion and decay lead to a dilution and finally the vanishing of the trace after some time. The *static floor field* S_{ij} does not change with time since it only takes into account the effects of the surroundings. Therefore it exists even without any pedestrians present. It allows to model, e.g., preferred areas, walls, and other obstacles. Figure 8 shows the static floor field used for the simulation of evacuations from a room with a single door. Its strength decreases with increasing distance from the door. Since the pedestrian prefers motion into the direction of larger fields, this is already sufficient to find the door.

Coupling constants control the relative influence of both fields. For a strong coupling to the static field, pedestrians will choose the shortest path to the exit. This corresponds to a “normal” situation. A strong coupling to the dynamic field implies a strong herding behavior where pedestrians try to follow the lead of others. This often happens in emergency situations.

The model uses a fully parallel update. Therefore conflicts can occur where different particles choose the same destination cell. Conflicts have been considered a technical problem for a long time, and usually the dynamics has been modified in order to avoid them. The simplest method is to update pedestrians sequentially instead of using a fully parallel dynamics. However, this leads to other problems, e.g., the identification of the relevant timescale. Therefore it has been suggested in Kirchner et al. (2003a) to take these conflicts seriously as an important part of the dynamics.

For the floor field model, it has been shown in Kirchner et al. (2003b) that the behavior becomes more realistic if not all conflicts are resolved in the sense that one of the pedestrians is allowed to move, whereas the others stay at their positions.



Modelling of Pedestrian and Evacuation Dynamics, Fig. 8 Left: Static floor field for the simulation of an evacuation from a large room with a single door. The door is located in the middle of the upper boundary, and the field strength is increasing with increasing intensity. Right: Snapshot of the dynamical floor field created by people leaving the room

Instead with probability μ , which is called friction parameter, the movement of *all* involved pedestrians is denied (Kirchner et al. 2003b) (see Fig. 9). This allows to describe clogging effects between the pedestrians in a much more detailed way. Although a local effect, it can have enormous influence on macroscopic quantities like flow and evacuation time.

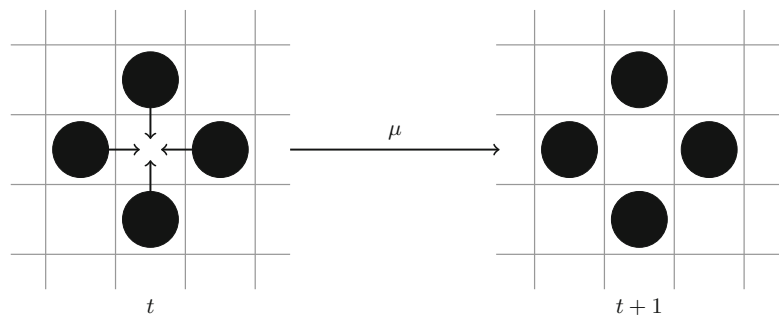
The original floor field model has been extended in various ways. An intrinsic problem of CA models is the limited spatial resolution which is a consequence of the discreteness in space. This makes it, e.g., impossible to represent a corridor of width 1 m by 40×40 (cm)² cells. To avoid this problem, models with smaller cells (e.g., 40×40 (cm)²) have been proposed (Kirchner et al. 2004). In this case pedestrians correspond to extended particles that occupy more than one cell (e.g., four cells).

The lattice structure introduces a spatial anisotropy. As a consequence straight motion in directions not parallel to the lattice axis is difficult to achieve. This can to some degree be cured by a proper choice of the transition probabilities (Burstedde et al. 2001; Schultz and Fricke 2010). For similar reasons in some models, agents are allowed to move further than just to nearest neighbor cells similar to the Blue-Adler model (Kirchner et al. 2004; Kretz 2007; Kretz and Schreckenberg 2006; Luo et al. 2018).

Most variants, however, have introduced other types of interactions. For counterflow situations the inclusion of anticipation effects (Nowak and Schadschneider 2012; Suma et al. 2011) leads to a more realistic behavior. Another interesting extension is the inclusion of game theory, e.g., in the

Modelling of Pedestrian and Evacuation Dynamics,

Fig. 9 Refused movement due to friction parameter μ



resolution of conflicts (Bouzat and Kuperman 2014) or the choice of exits in an evacuation (Lo et al. 2006).

Other CA Models

There are many other cellular automata models that have been proposed in the last 20 years. Most of them can be considered generalizations of the ASEP. One of the earliest is the Fukui-Ishibashi model (Fukui and Ishibashi 1999a, b) which is based on a two-dimensional variant of the ASEP. Originally it was used to study bidirectional motion in a long corridor. The motion is deterministic; only sidestepping to avoid oppositely moving pedestrian is stochastic. Various extensions and variations of the model have been proposed, e.g., an asymmetric variant (Muramatsu et al. 1999) where walkers prefer lane changes to the right, different update types (Fang et al. 2003), simultaneous (exchange) motion of pedestrians standing “face to face” (Jian et al. 2005), or the possibility of backstepping (Maniccam 2005). The influence of the shape of the particles has been investigated in Nagai and Nagatani (2006). Also other geometries (Muramatsu and Nagatani 2000b; Tajima and Nagatani 2002) and extensions to full two-dimensional motion have been studied in various modifications (Maniccam 2003, 2005; Muramatsu and Nagatani 2000a).

In the Gipps-Marksjö model (Gipps and Marksjö 1985), interactions between pedestrians are assumed to be repulsive, anticipating the idea of social forces (see section “Acceleration-Based Models”). To each cell a score is assigned based on its proximity to other pedestrians. This score represents the repulsive interactions, and the actual motion is then determined by the competition between these repulsion and the gain of approaching the destination. The pedestrian then selects the cell of its nine neighbors (Moore neighborhood) which leads to the maximum benefit which is given by the difference between the gain of moving closer to the destination and the cost of moving closer to other pedestrians. This leads to a deterministic model dynamics. Updating is done sequentially to avoid conflicts.

Related Approaches

Lattice-Gas Models

Conceptually, lattice-gas models are very similar to cellular automata models. For the modelling of pedestrian dynamics, the main difference is usually the position of the particles. In lattice-gas models the pedestrians are located on the vertices of the lattice, whereas they are located on the faces of the lattice in CA. Apart from this, the definition of the dynamics is analogous.

However, there are models of pedestrian dynamics which take inspiration from lattice-gas models that have been introduced for the description of classical hydrodynamics (Frisch et al. 1986), e.g., the model introduced by Marconi and Chopard (2002) which is based on a sequence of collision and propagation steps.

Semi-Discrete Models

In Yamamoto et al. (2007) an approach based on real-coded cellular automata (RCA) has been proposed. In contrast to a genuine CA, the velocities of the agents in a RCA are not discrete, and agents can move in any direction. Therefore the velocity is not given by the number of lattice sites the agent moves. The new position is determined from the non-integer velocity by using a rounding procedure. RCA models have the advantage of making the motion isotropic irrespective of the lattice structure.

In the Optimal Steps Model (Seitz and Köster 2012), the pedestrian movement takes place in continuous space and is discretized in a way that reflects the movement in steps. The model is built in order to achieve natural grid-free trajectories without introducing social force-type differential equations or complex steering behaviors.

Macroscopic Models

Fluid-Dynamic Models

Pedestrian dynamics has some obvious similarities with fluids. For example the motion around obstacles appears to follow “streamlines.” Therefore it is not surprising that, very much like for vehicular dynamics, the earliest models of

pedestrian dynamics took inspiration from hydrodynamics or gas-kinetic theory (Bellomo et al. 2012; Helbing 1992; Henderson 1974). Typically these macroscopic models are deterministic, force-based, and of low fidelity.

Motivated by the observation that the velocity distribution functions of crowds at low densities have a good agreement with the Maxwell-Boltzmann distribution (Henderson 1971), Henderson has developed a fluid-dynamic theory of pedestrian flow (Henderson 1974). Describing the interactions between the pedestrians as a collision process where the particles exchange momenta and energy, a homogeneous crowd can be described by the well-known kinetic theory of gases. However, the interpretation of the quantities is not entirely clear, e.g., what the analogues of pressure and temperature are in the context of pedestrian motion. Temperature could be identified with the velocity variance, which is related to the distribution of desired velocities, whereas the pressure expresses the desire to move against a force in a certain direction.

Hughes (2000, 2002, 2003) has derived a fluid-dynamical model which he calls “thinking fluid” based on a function measuring the discomfort of a pedestrian at a given density and the tendency to minimize the estimated travel time. It leads to a nonlinear system of partial differential equations involving an eikonal equation.

The applicability of classical hydrodynamic models is based on several conservation laws. The conservation of mass, corresponding to conservation of the total number of pedestrians, is expressed through a continuity equation of the form

$$\frac{\partial \rho(\mathbf{r}, t)}{\partial t} + \nabla \cdot \mathbf{J}(\mathbf{r}, t) = 0, \quad (10)$$

which connects the local density $\rho(\mathbf{r}, t)$ with the current $\mathbf{J}(\mathbf{r}, t)$. This equation can be generalized to include source and sink terms. However, the assumption of conservation of energy and momentum is not true for interactions between pedestrians which in general do not even satisfy Newton’s third law (“actio = reactio”). In Helbing (1992), several other differences to normal fluids

were pointed out, e.g., the anisotropy of interactions or the fact that pedestrians usually have an individual preferred direction of motion.

Gaskinetic Models

In Helbing (1992) a fluid-dynamical description was derived on the basis of a gaskinetic model which describes the system in terms of a density function $f(\mathbf{r}, \mathbf{v}, t)$. The dynamics of this function is determined by Boltzmann’s transport equation that describes its change for a given state as difference of inflow and outflow due to binary collisions.

An important new aspect in pedestrian dynamics is the existence of desired directions of motion which allows to distinguish different groups μ of particles. The corresponding densities f_μ change in time due to four different effects:

1. A relaxation term with characteristic time τ describes tendency of pedestrians to approach their intended velocities.
2. The interaction between pedestrians is modelled by a Stosszahlansatz as in the Boltzmann equation. Here pair interactions between types μ and ν occur with a total rate that is proportional to the densities f_μ and f_ν .
3. Pedestrians are allowed to change from type μ to ν which, e.g., accounts for turning left or right at a crossing.
4. Additional gain and loss terms allow to model entrances and exits where pedestrians can enter or leave the system.

The resulting fluid-dynamic equations are similar to that of ordinary fluids although, due to the different types of pedestrians, one actually obtains a set of coupled equations describing several interacting fluids. Equilibrium is approached through the tendency to walk with the intended velocity, not through interactions as in ordinary fluids. Momentum and energy are not conserved in pedestrian motion, but the relaxation toward the intended velocity describes a tendency to restore these quantities. Unsurprisingly for a macroscopic approach, the gas-kinetic models have problems at low densities. For further discussion, see, e.g., Helbing (1992).

Network and Queuing Models

Another class of models is based on ideas from queuing theory. Typically rooms are represented as nodes in the queuing network, and links correspond to doors. In microscopic approaches each agent chooses a new node, e.g., according to some probability (Lovas 1994). In Treuille et al. (2006), a fluid-dynamic model is transformed into a particle system in a network. Some explicit particle-to-particle interactions are considered, for instance, to enforce a minimum distance between pedestrians.

Queuing models are generally simple to implement and fast to simulate. They are particularly suitable to large-scale simulation of pedestrian movement. For instance, the queue simulation model introduced by Lämmel et al. is used for the evacuation of the city of Padang (Indonesian) in case of a tsunami warning (Lämmel et al. 2009). The affected population comprised approximately 330,000 individuals, and the overall network consists of 6289 nodes, 16,978 links, and more than 900,000 cells in an area of roughly $7 \times 4 \text{ km}^2$. Few minutes (approximately 400 s) were necessary in 2011 to simulate the full evacuation on a standard computer.

Lastly, multi-scale models on network may be mentioned. Multi-scale (or hybrid) models combine models formulated at different levels of aggregation yielding in a compromise between

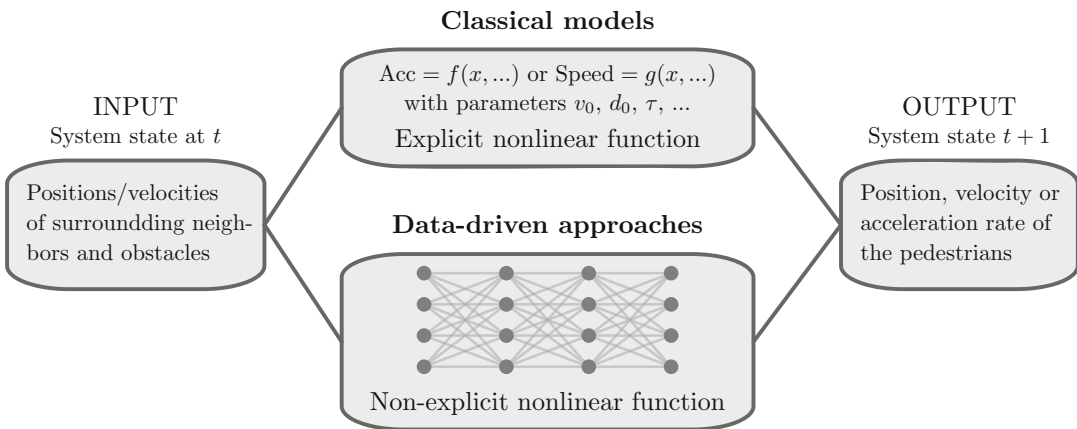
computational complexity and accuracy (Crociani et al. 2016; Lämmel et al. 2014). Ensuring consistency across levels, and especially transitions from a level to another, both in terms of parameter calibration and in terms of state variables, makes the use of such models however relatively difficult.

Future Directions and Open Problems

Data-Driven Modelling

The increase of real and experimental database and computational capacity of the computers makes nowadays possible the development of machine learning approaches for prediction of pedestrian movement. Oppositely to physical models, data-driven approaches are deliberately complex and have a high number of parameters with no possible physical interpretations (see Fig. 10). Their calibration (or *training*) generally requires a very large amount of data. Accurately trained, the very high plasticity of the prediction models allows in theory to describe any type of patterns.

The methodological approach with data-driven approaches consists in partitioning the dataset in training, testing, and even predicting subsets. Bootstrap techniques are used to evaluate the quality (precision) of the prediction (Kohavi



Modelling of Pedestrian and Evacuation Dynamics, Fig. 10 Minimalistic illustrative scheme for the distinction between classical models calibrated by few

meaningful parameters, and data-driven modelling, for which the number of parameters is deliberately large

1995). Data-driven modelling approach is used for predictions of pedestrian movements, notably in complex geometries (Das et al. 2015), for the motion planning of robots moving in a crowd (Chen et al. 2018), or for modelling of walker pose (Fragkiadaki et al. 2015). Simplest approaches are based on artificial neural networks (see, e.g., Das et al. (2015) and Ma et al. (2016)), while the most sophisticated prediction models lie in Long-Short-Term Memory networks (Alahi et al. 2016) or deep reinforcement learning techniques (Chen et al. 2017).

The predictions with machine learning approaches are generally more precise than those of physical models; however their robustness for different types of facilities and walking situations remains to be demonstrated. However, the development of data-driven approaches is expected to grow during the next years. This can be for prediction of pedestrian movement in complex facilities, and potentially also of evacuations, as well as for motion planning of robots moving in a crowded environment. Yet the calibration (training) and validation of machine learning models are difficult tasks. One of the reasons lies in the fact that the amount data necessary for calibration depends on the complexity of the modelling approach. A too complex approach, for instance, leads to overfitting of the data used for the calibration and bad prediction of new data.

Validation and Calibration

All models should be validated before they are used in practical applications. If quantitative predictions are required, e.g., estimates of evacuation times, the models also should be calibrated properly. So far there seems to be no real consensus how this can be achieved.

As discussed in the companion article (Boltes et al. 2018), empirically several interesting collective phenomena are observed in crowd motion, e.g., clogging at bottlenecks, stop-and-go motion, dynamical lane formation in counterflow, etc. Currently the qualitative reproduction of these effects is often used as criterion for the validity of a model.

The main problem with calibration is the lack of reliable empirical data (Seyfried and

Schadschneider 2009). As described in Boltes et al. (2018) even for fundamental diagram, i.e., the density dependence of the flow, currently no consensus exists. Published data for the fundamental diagram – and other quantities as well – differ substantially. The origin of these discrepancies is – up to now – not clear and one of the most important areas of research. The situation is thus very different from that in vehicular traffic. Here a basic consensus about the quantitative dependencies exists, although there is still some debate about the interpretation of the data.

The lack of generally accepted quantitative empirical results causes problems for the calibration of models. In highway traffic the empirical fundamental diagram is the main source for calibration. However, the published fundamental diagrams of pedestrian streams differ quantitatively so much that they cannot be considered to yield reliably calibrated models. Therefore the development of a proper procedure for model calibration remains one of the most important issues of pedestrian dynamics (Seyfried and Schadschneider 2009).

As outlined in the [introduction](#), different research areas pursuing different objectives contribute to the modelling of pedestrian dynamics. This diversity makes it difficult to establish a consensus with respect to minimal standards or a general accepted suite of benchmarks for models.

Route Choice and Navigation

While most models focus on the operational level, only few works exist to model route choices and navigation of agents on the tactical or strategic level; see section “[Classification of Models](#).”

For application of evacuation, the shortest path is a standard approach (Dijkstra 1958). A more advanced method is the representation of routes according to graphs (see Höcker (2010) and Höcker et al. (2010). To optimize the travel time by considering waiting times, e.g., for the selection of doors or alternative routes, see Kemloh Wagoum et al. (2012), Kretz (2009) and Kretz et al. 2015).

Several approaches exist taking human wayfinding abilities into account. The impact of discrete choices at the knots of the network was modelled by Løvs (1998). Herding as a strategy of

unfamiliar agents was used in Ehtamo et al. (2010) by introducing game theoretical approaches to model imitation of decisions of other agents. In Kneidl (2013)) spatial inaccuracies concerning metrics and direction were used for wayfinding. This approach was extended by Kielar et al. (2016b) to agents with different knowledge. To enable agents finding an exit without any knowledge, search strategies like random search and nearest room/door heuristic were used in von Sivers et al. (2016). Models for the perception of signage including the dependency on the signs' properties, the distances, the observation angles to signs, or lighting conditions could be found in Filippidis (2006), Nassar (2011), and Xie et al. (2007).

An approach to model partial knowledge integrating visibility graphs with a cognitive map approach including landmarks, generalized knowledge like the functionality of corridors, and stairs and signage can be found in Andresen et al. (2018).

Simulation Frameworks

For application, several models are implemented in different software. Besides commercial software several open-source projects like Menge (Curtis et al. 2016), FDS + Evac (Korhonen et al. 2010), VADERE (Köster et al. 2017), or JuPedSim (Kemloh Wagoum et al. 2015) are available. These frameworks enable the user to define the number and initial positions of pedestrians, the parameter of the models, and the spatial layout of the buildings with the definition of usable exits. Kielar et al. (2016a) give an extensive overview of the existing pedestrian simulation frameworks. Meanwhile they introduce MomenTUMv2, a generic, modular, and flexible framework for agent-based simulation that enables users to couple different models in an easy way.

Bibliography

Primary Literature

Alahi A, Goel K, Ramanathan V, Robicquet A, Fei-Fei L, Savarese S (2016) Social lstm: human trajectory prediction in crowded spaces. In: 2016 I.E. conference on

- computer vision and pattern recognition (CVPR), p 961971
- Ali S, Nishino K, Manocha D, Shah M (eds) (2013) Modeling, simulation and visual analysis of crowds a multidisciplinary perspective. Springer, New York
- Andresen E, Chraibi M, Seyfried A (2018) A representation of partial spatial knowledge: a cognitive map approach for evacuation simulations. *Transportmetrica/A* n/a:1–34. online first
- Bellomo N, Piccoli B, Tosin A (2012) Modeling crowd dynamics from a complex system viewpoint. *Math Models Methods Appl Sci* 22(Suppl 2):123004
- Ben-Jacob E (1997) From snowflake formation to growth of bacterial colonies. Part II. Cooperative formation of complex colonial patterns. *Contemp Phys* 38:205
- Biham O, Middleton AA, Levine D (1992) Self-organization and a dynamical transition in traffic-flow models. *Phys Rev A* 46:R6124
- Blue VJ, Adler JL (2000) Cellular automata microsimulation of bi-directional pedestrian flows. *J Transp Res Board* 1678:135–141
- Blue VJ, Adler JL (2002) Flow capacities from cellular automata modeling of proportional splits of pedestrians by direction. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg
- Blythe R, Evans MR (2007) Nonequilibrium steady states of matrix product form: a solver's guide. *J Phys A* 40: R333
- Boltes M, Zhang J, Tordeux A, Schadschneider A, Seyfried A (2018) Pedestrian and evacuation dynamics: empirical results. In: *Encyclopedia of complexity and system science*. Springer
- Bouzat S, Kuperman MN (2014) Game theory in models of pedestrian room evacuation. *Phys Rev E* 89:032806
- Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A* 295:507–525
- Burstedde C, Kirchner A, Klauck K, Schadschneider A, Zittartz J (2002) Cellular automaton approach to pedestrian dynamics – applications. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 87–98
- Chen Y, Everett M, Liu M, How JP (2017) Socially aware motion planning with deep reinforcement learning. In: *Proceedings of the 2017 I.E. international conference on intelligent robots and systems (IROS)*, Vancouver, British Columbia, Canada, pp 1343–1350
- Chen X, Treiber M, Kanagaraj V, Li H (2018) Social force models for pedestrian traffic – state of the art. *Transp Rev* 38(5):625–653
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329(4–6):199–329
- Chraibi M, Seyfried A, Schadschneider A (2010) Generalized centrifugal force model for pedestrian dynamics. *Phys Rev E* 82:046111
- Chrastil ER, Warren WH (2015) Active and passive spatial learning in human navigation: Acquisition of graph knowledge. *J Exp Psychol Learn Mem Cogn* 41:1162

- Crociani L, Lämmel G, Vizzari G (2016) Multi-scale simulation for crowd management: A case study in an urban scenario. In: Autonomous agents and multiagent systems. Springer International Publishing, Cham pp 147–162
- Curtis S, Best A, Manocha D (2016) Menge: a modular framework for simulating crowd movement. *Collective Dyn* 1(0):1–40
- Daamen W (2004) Modelling passenger flows in public transport facilities. PhD thesis, Technical University of Delft
- Das P, Parida M, Katiyar VK (2015) Analysis of interrelationship between pedestrian flow parameters using artificial neural network. *J Mod Transp* 23(4):298–309
- Derrida B (1998) An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Phys Rep* 301:65
- Dietrich F, Köster G (2014) Gradient navigation model for pedestrian dynamics. *Phys Rev E* 89:062801
- Dijkstra EW (1958) A note on two problems in connexion with graphs. *Numer Math* 1(1):269271
- Ehtamo H, Heliövaara S, Korhonen T, Hostikka S (2010) Game theoretic best-response dynamics for evacuees' exit selection. *Adv Complex Syst* 13(01):113–134
- El Yacoubi S, Chopard B, Bandini S (eds) (2006) Cellular automata -7th international conference on cellular automata for research and industry, ACRI 2006, Perpignan, France, 2006. Springer, Berlin/Heidelberg/New York
- Filippidis L (2006) Representing the influence of signage on evacuation behavior within an evacuation model. *J Fire Prot Eng* 16(1):37–73
- Fiorini P, Shiller Z (1998) Motion planning in dynamic environments using velocity obstacles. *Int J Robot Res* 17:760772
- Fragkiadaki K, Levine S, Felsen P, Malik J. Recurrent network models for human dynamics. In: Proceedings of the 2015 I.E. international conference on computer vision (ICCV), ICCV '15, Washington, DC, USA, 2015. IEEE Computer Society, pp 4346–4354
- Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the Navier-Stokes equation. *Phys Rev Lett* 56:1505
- Fukui M, Ishibashi Y (1999a) Jamming transition in cellular automaton models for pedestrians on passageway. *J Phys Soc Jpn* 68:3738
- Fukui M, Ishibashi Y (1999b) Self-organized phase transitions in cellular automaton models for pedestrians. *J Phys Soc Jpn* 68:2861
- Galea ER (ed) (2003) Pedestrian and evacuation dynamics 2003. CMS Press, London
- Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. *Math Comput Simul* 27:95–105
- Guy S, Chhugani J, Kim C, Satish S, Lin M, Manocha D, Dubey P (2009) Clearpath: highly parallel collision avoidance for multi-agent simulation. In: Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation 2009, New Orleans, LA, USA, pp 177–187
- Guy S, Chhugani J, Curtis S, Dubey P, Lin M, D Manocha. (2010a) PLEdestrians: a least-effort approach to crowd simulation. In: Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation. Eurographics Association 2010, Madrid, Spain, pp 119–128
- Guy S, Lin M, Manocha D (2010b) Modeling collision avoidance behavior for virtual humans. In: Proceedings of the 9th international conference on autonomous agents and multiagent systems, Toronto, ON, Canada, vol 2, pp 575–582
- Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. *Complex Syst* 6:391–415
- Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51:4282–4286
- Helbing D, Farkas I, Molnár P, Vicsek T (2002) Simulation of pedestrian crowds in normal and evacuation situations. In: Schreckenberg M, Sharma SD (eds) Pedestrian and evacuation dynamics. Springer, Berlin/Heidelberg, pp 21–58
- Henderson LF (1971) The statistics of crowd fluids. *Nature* 229:381–383
- Henderson LF (1974) On the fluid mechanics of human crowd motion. *Transp Res* 8:509–515
- Hirai K, Tarui K (1975) A simulation of the behavior of a crowd in panic. In: Proceedings of the 1975 international conference on cybernetics and society, San Francisco, pp 409–411
- Höcker M (2010) Modellierung und Simulation von Fußgängerverkehr – Entwicklung mathematischer Ansätze für Soziale Kräfte und Navigationsgraphen. PhD thesis, Fakultät für Bauingenieurwesen und Geodäsie der Gottfried Wilhelm Leibniz Universität Hannover
- Höcker M, Berkahn V, Kneidl A, Borrmann A, Klein W (2010) Graph-based approaches for simulating pedestrian dynamics in building models. In: Menze E, Scherer R (eds) ECPPM 2010 – eWork and eBusiness in Architecture, Engineering and Construction, pp 389–394
- Hoogendoorn SP, Bovy PHL, Daamen W (2002) Microscopic pedestrian wayfinding and dynamics modelling. In: Pedestrian and evacuation dynamics, pp 123–154
- Hughes RL (2000) The flow of large crowds of pedestrians. *Math Comput Simul* 53:367–370
- Hughes RL (2002) A continuum theory for the flow of pedestrians. *Transp Res B* 36:507–535
- Hughes RL (2003) The flow of human crowds. *Annu Rev Fluid Mech* 35:169
- Jian L, Lizhong Y, Daoling Z (2005) Simulation of bi-direction pedestrian movement in corridor. *Physica A* 354:619
- Kemloh Wagoum AU, Seyfried A, Holl S (2012) Modeling the dynamic route choice of pedestrians to assess the criticality of building evacuation. *Adv Complex Syst* 15:1250029
- Kemloh Wagoum AU, Chraibi M, Zhang J (2015) Jupedsim: an open framework for simulating and analyzing the dynamics of pedestrians. In: 3rd Conference of Transportation Research Group of India
- Kielar P, Biedermann D, Borrmann A (2016a) Momentumv2: a modular, extensible, and generic agent-based pedestrian behavior simulation framework. Technical report, Technical University Munich

- Kielar PM, Biedermann DH, Kneidl A, Borrmann A (2016b) A unified pedestrian routing model combining multiple graph-based navigation methods. In: Knoop VL, Daamen W (eds) *Traffic and granular flow '15*. Springer International Publishing, Cham, pp 241–248
- Kim S, Stephen G, Guy S, Liu W, Wilkie D, Lau R, Lin M, Manocha D (2014) Brvo: predicting pedestrian trajectories using velocity-space reasoning. *Int J Rob Res* 34:201–217
- Kirchner A, Schadschneider A (2002) Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. *Physica A* 312:260–276
- Kirchner A, Namazi A, Nishinari K, Schadschneider A (2003a) Role of conflicts in the floor field cellular automaton model for pedestrian dynamics. In: Galea (2003), p 51
- Kirchner A, Nishinari K, Schadschneider A (2003b) Friction effects and clogging in a cellular automaton model for pedestrian dynamics. *Phys Rev E* 67:056122
- Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization effects and the influence of walking speed in cellular automata models for pedestrian dynamics. *J Stat Mech* 2004(10):P10011
- Klüpfel H, Meyer-König T, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passenger ships. In: Bandini S, Worsch T (eds) *Theory and practical issues on cellular automata*. Springer, Berlin/Heidelberg
- Kneidl A (2013) *Methoden zur Abbildung menschlichen Navigationsverhaltens bei der Modellierung von Fußgängerströmen*. PhD thesis, Technische Universität München
- Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th international joint conference on artificial intelligence – volume 2, IJCAI'95*, San Francisco, CA, USA. Morgan Kaufmann Publishers, pp 1137–1143
- Korhonen T, Hostikka S, Heliövaara S, Ehtamo H (2010) Fds+evac: an agent based fire evacuation model. In: Klingsch WWF, Rogsch C, Schadschneider A, Schreckenberg M (eds) *Pedestrian and evacuation dynamics 2008*. Springer, Berlin/Heidelberg, pp 109–120
- Köster G, Tremel F, Gödel M (2013) Avoiding numerical pitfalls in social force models. *Phys Rev E* 87:063305
- Kretz T (2007) *Pedestrian traffic – simulation and experiments*. PhD thesis, Universität Duisburg-Essen
- Kretz T (2009) *Pedestrian traffic: On the quickest path*. *J Stat Mech: Theory Exp* 2009(03):P03012
- Kretz T (2015) On oscillations in the social force model. *Physica A* 438:272–285
- Kretz T, Schreckenberg M (2006) The F.A.S.T.-model. *Lecture Notes in Computer Science* 4173:712
- Kretz T, Lehmann K, Hofsaß I (2015) Pedestrian route choice by iterated equilibrium search. In: Chraïbi M, Boltes M, Schadschneider A, Seyfried A (eds) *Traffic and granular flow '13*. Springer International Publishing, Cham, pp 47–54
- Lämmel G, Klüpfel H, Nagel K (2009) Chapter 11. The MATSim network flow model for traffic simulation adapted to large-scale emergency egress and an application to the evacuation of the Indonesian City of Padang in Case of a Tsunami Warning. in Harry Timmermans (ed.) *Pedestrian Behavior* Emerald Publishing Limited, pp 245–265
- Lämmel G, Seyfried A, and B. Steffen (2014) Large-scale and microscopic: a fast simulation approach for urban areas. In: *Transportation Research Board Annual Meeting*, Washington DC, USA, pp 3814–3890
- Lo SM, Huang HC, Wang P, Yuen KK (2006) A game theory based exit selection model for evacuation. *Fire Saf J* 41:364
- Lovas GG (1994) Modeling and simulation of pedestrian traffic flow. *Transp Res B* 28V:429
- Løvs GG (1998) Models of wayfinding in emergency evacuations. *Eur J Oper Res* 105(3):371–389
- Luo L, Fu Z, Cheng H, Yang L (2018) Update schemes of multi-velocity floor field cellular automaton for pedestrian dynamics. *Physica A* 491:946
- Ma Y, Lee EWM, Yuen RKK (2016) An artificial intelligence-based approach for simulating pedestrian movement. *IEEE Trans Intell Transp Syst* 17(11):3159–3170
- Maniccam S (2003) Traffic jamming on hexagonal lattice. *Physica A* 321:653
- Maniccam S (2005) Effects of back step and update rule on congestion of mobile objects. *Physica A* 346:631
- Marconi S, Chopard B (2002) A multiparticle lattice gas automata model for a crowd. *Lecture Notes in Computer Science*, 2493, p 231, Springer, Berlin, Heidelberg
- Maury B, Venel J (2011) A discrete contact model for crowd motion. *ESAIM: Math Model Numer Anal* 45:145–168
- Maury B, Venel J, Olivier A-H, Donikian S (2008) A mathematical framework for a crowd motion model. *Comptes Rendus de l'Académie des Sciences – Series I* 346:1245–1250
- Mordvintsev A, Krzhizhanovskaya V, Lees M, Sloat P (2014) Simulation of city evacuation coupled to flood dynamics. In: *Pedestrian and evacuation dynamics 2012*. Springer, Cham, pp 485–499
- Moussaïd M, Helbing D, Garnier S, Johansson A, Combe M, Theraulaz G (2009) Experimental study of the behavioural mechanisms underlying self-organization in human crowds. *Proc R Soc B* 276(1668):2755–2762
- Muramatsu M, Nagatani T (2000a) Jamming transition in two-dimensional pedestrian traffic. *Physica A* 275:281–291
- Muramatsu M, Nagatani T (2000b) Jamming transition of pedestrian traffic at crossing with open boundary conditions. *Physica A* 286:377–390
- Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. *Physica A* 267:487–498
- Nagai R, Nagatani T (2006) Jamming transition in counter flow of slender particles on square lattice. *Physica A* 366:503
- Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I France* 2:2221
- Nassar K (2011) Sign visibility for pedestrians assessed with agent-based simulation. *Transp Res Rec J Transp Res Board* 2264:18–26

- Nowak S, Schadschneider A (2012) Quantitative analysis of pedestrian counter-flow in a cellular automaton model. *Phys Rev E* 85:066128
- Ondrej J, Pettré J (2010) A synthetic-vision based steering approach for crowd simulation. *ACM Trans Graph* 29:123
- Paris S, Pettré J, Donikian S (2007) Pedestrian reactive navigation for crowd simulation: a predictive approach. *Comput Graph Forum* 26:665–674
- Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24(3):274–281
- Rickert M, Nagel K, Schreckenberg M, Latour A (1996) Two lane traffic simulations using cellular automata. *Physica A* 231:534
- Rio KW, Rhea CK, Warren WH (2014) Follow the leader: visual control of speed in pedestrian following. *J Vis* 14(2):4
- Schadschneider A (2002) Cellular automaton approach to pedestrian dynamics -theory. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 75–86
- Schreckenberg M, Sharma SD (eds) (2002) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg
- Schultz M, Fricke H (2010) Stochastic transition model for discrete agent movements. *Lecture Notes in Computer Science* 6350:506
- Schütz GM (2001) Exactly solvable models for many-body systems. In: Domb C, Lebowitz JL (eds) *Phase transitions and critical phenomena*, vol 19. Academic, London
- Seitz M, Köster G (2012) Natural discretization of pedestrian movement in continuous space. *Phys Rev E* 86:046108
- Seitz M, Dietrich F, Köster G, Bungartz H-J (2016) The superposition principle: a conceptual perspective on pedestrian stream simulations. *Collective Dyn* 1:1–19
- Seyfried A, Schadschneider A (2009) Validation of ca models of pedestrian dynamics with fundamental diagrams. *Cybern Syst* 40:367
- Sieben A, Schumann J, Seyfried A (2017) Collective phenomena in crowds -where pedestrian dynamics need social psychology. *PLoS One* 12:1–19
- Suma Y, Yanagisawa D, Nishinari K (2011) Anticipation effect in pedestrian dynamics: modeling and experiments. *Physica A* 391:248
- Tajima Y, Nagatani T (2002) Clogging transition of pedestrian flow in t-shaped channel. *Physica A* 303:239–250
- Templeton A, Drury J, Philippides A (2015) From mindless masses to small groups: conceptualizing collective behavior in crowd modeling. *Rev Gen Psychol* 19(3):215–229
- Köster G et al. (2017). *VADERE*, <http://www.vadere.org/>
- Tordeux A, Schadschneider A (2016) White and relaxed noises in optimal velocity models for pedestrian flow with stop-and-go waves. *J Phys A Math Theor* 49:185101
- Tordeux A, Chraïbi M, Seyfried A (2016) Collision-free speed model for pedestrian dynamics. In: *Traffic and granular flow '15*, Springer, Cham pp 225–232
- Treuille A, Cooper S, Popovic Z (2006) Continuum crowds. *ACM Trans Graph/SIGGRAPH* 25(3): 1160–1168
- van den Berg J, Lin M, Manocha D (2008) Reciprocal velocity obstacles for realtime multi-agent navigation. In: 2008 I.E. international conference on robotics and automation, Pasadena, CA, USA, pp 1928–1935
- van den Berg J, Guy S, Lin M, Manocha D (2011) Reciprocal n-body collision avoidance. *Robotics research: the 14th international symposium ISRR*, Lucerne, Switzerland, pp 3–19
- von Sivers I, Seitz MJ, Köster G (2016) How do people search: a modelling perspective. *Lecture notes in computer science*. 9574:487–496
- Weidmann U. *Transporttechnik der Fußgänger – Transporttechnische Eigenschaften des Fußgängerverkehrs (Literaturauswertung)*. Schriftenreihe des IVT 90, ETH Zürich, 3 1993. Zweite, ergänzte Auflage (in German)
- Fang W, Yang L, Fang W (2003) Simulation of bi-directional pedestrian movement using a cellular automata model. *Physica A* 321:633–640
- Xie H, Filippidis L, Gwynne S, Galea ER, Blackshields D, Lawrence PJ (2007) Signage legibility distances as a function of observation angle. *J Fire Prot Eng* 17(1):41–64
- Yamamoto K, Kokubo S, Nishinari K (2007) Simulation for pedestrian dynamics by real-coded cellular automata (RCA). *Physica A* 379:654

Books and Reviews

- Chopard B, Droz M (1998) *Cellular automaton modeling of physical systems*. Cambridge University Press, Cambridge
- Knoop VL, Daamen W (eds) (2016) *Traffic and granular flow '15*. Springer, Cham (see also previous issues of this conference series)
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and granular flow '05*. Springer, Berlin
- Schadschneider A, Chowdhury D, Nishinari K (2010) *Stochastic transport in complex systems: from molecules to vehicles*. Elsevier, Amsterdam
- Schreckenberg M, Sharma SD (eds) (2002) *Pedestrian and evacuation dynamics*. Springer, Berlin
- Song W, Ma J, Fu L (eds). *Pedestrian and evacuation dynamics 2016*. Available from <http://collective-dynamics.eu/index.php/cod/article/view/A11>
- Still K (2013) *Introduction to crowd science*. CRC Press, Boca Raton
- Thalmann D, Raupp Musse S (2007) *Crowd simulation*. Springer, London
- Timmermans H (ed) (2009) *Pedestrian behavior – models, data collection and applications*. Emerald, Bingley
- Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2007) *Pedestrian and evacuation dynamics '05*. Springer, Berlin

Empirical Results of Pedestrian and Evacuation Dynamics

Maik Boltes¹, Jun Zhang², Antoine Tordeux³,
Andreas Schadschneider⁶ and Armin Seyfried^{4,5}

¹Institute for Advanced Simulation,
Forschungszentrum Jülich, Jülich, Germany

²State Key Laboratory of Fire Science, University
of Science and Technology of China, Hefei, China

³School of Mechanical Engineering and Safety
Engineering, University of Wuppertal,
Wuppertal, Germany

⁴Institute for Advanced Simulation,
Forschungszentrum Jülich GmbH, Jülich,
Germany

⁵School of Architecture and Civil Engineering,
University of Wuppertal, Wuppertal, Germany

⁶Institut für Theoretische Physik, Universität zu
Köln, Köln, Germany

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Definition of Observables
Collective Effects
Fundamental Diagram
Evacuations and Crowd Disasters
Controlled Experiments
Conclusions
Future Directions
Bibliography

Glossary

Bottleneck A bottleneck is in general a part of facility limiting pedestrian flows. This can be, for example, a door, a narrowing in a corridor, or stairs, i.e., locations of reduced capacity. At bottlenecks jamming occurs if the inflow is higher than the capacity.

Capacity The maximal flow rate supported by a facility is called “capacity.”

Crowd disaster Crowd disaster is an accident in which the specific behavior of the crowd is a relevant factor, e.g., through competitive and nonadaptive behavior. In the media, it is often called “panic” which is a scientifically not proven concept in crowd dynamics and should thus be avoided.

Crowd A large group of pedestrians who have gathered together. Depending on the perspective, more specific definitions exist.

Evacuation Evacuation is the movement of persons from a dangerous place due to the threat or occurrence of a disastrous event. In normal situations this is called “egress” instead.

Flow rate The flow rate J is a measure for the throughput and describes the performance of a pedestrian facility. It is defined as the number of persons passing a cross section per unit time. The common unit of flow is “persons per second.” Often the specific flow is used which is the flow rate per unit width.

Fundamental diagram In traffic engineering (and physics), the fundamental diagram describes the density dependence of the flow: $J(\rho)$. Due to the hydrodynamic relation $J = \rho vb$, equivalent representations used frequently are $v(\rho)$ and $v(J)$. The fundamental diagram is probably the most important quantitative characterization of traffic systems.

Hydrodynamic relation The hydrodynamic relation $J = \rho v$ connects the flow rate J with the density ρ and the average velocity v in a one-dimensional particle stream. For motion in two dimensions, it is given by $J = \rho vb$ where b is the width.

Lane formation In bidirectional flows, lanes in which all pedestrians move in the same direction are often formed dynamically. The number of lanes is usually not constant in time and can be different for the different directions.

Macroscopic models Macroscopic models do not distinguish individuals. The description is based on aggregate quantities, e.g., appropriate

densities. Typical models belonging to this class are fluid-dynamic approaches.

Microscopic models Microscopic models represent each pedestrian separately with individual properties like walking velocity or route choice behavior and the interactions between them.

Pedestrian A pedestrian is a person travelling on foot. In this entry, other characterizations are used, depending on the context, e.g., agent or particle.

Trajectory A trajectory of a pedestrian is its path in time. For controlled experiments this is most often derived from the motion of the top of the head including swaying and bobbing. The individual trajectory can be enriched with information changing over time (e.g., head or shoulder orientation) or static information (e.g., age or gender).

Definition of the Subject and Its Importance

Today, there are many occasions where a large number of people gather in a relatively limited space. Very large events related to sports, entertainment, culture, and religion are held all over the world on a regular basis. Office buildings and apartment houses become much larger and more complex. This brings about serious safety issues for the participants and the organizers who have to prepare for any case of emergency or critical situation.

To reduce the risk of injuries or even fatalities in large crowds, it is important to understand the basic mechanism that governs crowd motion or crowd behavior in general. This requires reliable empirical data, both qualitative and quantitative, which can be analyzed to uncover the basic underlying principles. This then allows the development of realistic models which in turn can be validated and tested against the empirical data.

Despite the obvious importance of empirical data, it is surprising that no consensus about the correct quantitative description exists even for the simplest scenarios. This is quite different from the situations in vehicular traffic where vast datasets obtained in an automated way from detectors on highways exist. However, this is not so easy to be

transferred to pedestrian motion. The underlying physical principles of highway detectors (e.g., magnetic induction by moving metallic objects) cannot be applied here. Furthermore the motion of vehicles is usually (quasi-)one-dimensional, whereas pedestrian motion is a mixture of two-dimensional motions in different directions. This makes it difficult to use the analogue of floating-car data which is currently one of the standard techniques for obtaining data for vehicular traffic (Kerner 2017). Methods based on GPS measurements do not have a sufficient accuracy of the order of 1 cm which would be required to determine trajectories of pedestrians. Therefore new techniques have to be developed.

Introduction

One of the main goals of pedestrian science is to enhance safety in public places. The awareness that the reasonable setting of emergency exits is one of the most important factors to ensure the safety of occupants in buildings can be traced more than 100 years. However, the disasters due to the fires in the Ringtheater in Vienna and the urban theater in Nizza at 1881 with several hundred fatalities lead to a rethinking of the safety in buildings (Dieckmann 1911). Firstly it was tried to improve the safety by using nonflammable building materials. But the disaster at the Iroquois Theater in Chicago with more than 500 fatalities, where only the decoration is burned, caused a rethinking. It was a starting point for studying the influences of emergency exits and thus the dynamics of pedestrian streams (Dieckmann 1911; Fischer 1933).

In this context evacuation dynamics plays a special role. In general, evacuation is the egress from an area, a building, or vessel to a relatively safe place due to a potential or actual threat. The dynamics can be quite complex because of the large number of people and their interactions, external factors like fire, complex building geometries, etc. Evacuation dynamics has to be described and understood on different levels: physical, physiological, psychological, and social. Its investigation is difficult and involves many research areas and disciplines.

Due to joint and interdisciplinary efforts, much progress in the understanding of pedestrian and evacuation dynamics has been made in recent years. The origin of the apparent complexity lies in the fact that one is concerned with a many-“particle” system with complex interactions that are not fully understood. Therefore empirical investigations are required for a better understanding of the underlying processes. One of the new methods is laboratory experiments under controlled conditions, which can be repeated all over the world and has become more and more important. Using up to the order of 1000 participants allows to study subtle effects like the influence of the cultural background on the important quantities in traffic science, the flow-density relation, or the fundamental diagram.

In this entry we focus on empirical studies on pedestrian and evacuation dynamics but not modelling approaches which are reviewed in a companion article (Chraïbi et al. [n.d.](#)). Empirical results are important for the design, validation, and calibration of models. As we will show here, not for all situations, there is currently a consensus on the relevant empirical data. The origin of the discrepancies among different observations is not always clear. The resulting ambiguity involves serious problems in testing modelling approaches quantitatively.

The entry is organized as follows. In section “[Definition of Observables](#),” we introduce the definitions of the main observable that characterize the properties of pedestrian streams as well as discussing the problems in their determination. Section “[Collective Effects](#)” provides an overview of the main collective effects that have been found in systems of pedestrians. Probably the most important quantitative characterization of traffic systems is the fundamental diagram, which is described in section “[Fundamental Diagram](#)” for different geometries. In section “[Evacuations and Crowd Disasters](#),” observations from crowd disasters are discussed. These help to improve safety, e.g., by optimizing evacuation procedures. In section “[Controlled Experiments](#),” we describe how controlled laboratory experiments can contribute to a better understanding of pedestrian dynamics.

Definition of Observables

Introduction

Pedestrians are three-dimensional objects, and a complete description of their highly developed and intricate motion sequence is rather difficult. To extract transport characteristics of a system with many pedestrians, the motion is usually treated in two-dimensional space by considering the vertical projection of the body.

In the following sections, we review the present knowledge of empirical results. These are relevant not only as basis for the development of models but also for applications like safety studies and legal regulations.

We start with the phenomenological description of collective effects, some of which are known from daily experience and serve as benchmark tests for any kind of modelling approaches. Any model that is not able to reproduce these effects is missing some essential part of the dynamics. Next the foundations of a quantitative description are laid by introducing the fundamental observables of pedestrian dynamics. Difficulties arise from different definitions and inhomogeneities in space and time. Then pedestrian dynamics in several simple scenarios (corridor, stairs, etc.) is discussed. Surprisingly even for these simple cases, no consensus about the basic quantitative properties exists. Finally, more complex scenarios are discussed which are combinations of the simpler elements.

Before we review experimental studies, the commonly used observables are introduced.

Flow Rate

The flow rate J of a pedestrian stream gives the number of pedestrians crossing a fixed location of a facility per unit of time. Usually it is taken as a scalar quantity since only the flow normal to some cross section is considered. There are various methods to measure the flow rate. The most natural approach is to determine the times t_i at which pedestrians passed a fixed measurement location. The time gaps $\Delta t_i = t_{i+1} - t_i$ between two consecutive pedestrians i and $i+1$ are directly related to the flow rate.

$$J = \frac{1}{\langle \Delta t \rangle} \quad \text{with} \quad \langle \Delta t \rangle = \frac{1}{N} \sum_{i=1}^N (t_{i+1} - t_i) = \frac{t_{N+1} - t_1}{N}. \quad (1)$$

Such a definition allows to estimate the flow rate for a given number of observed pedestrians. Similarly, estimation of the flow for a given time interval can be obtained thanks to the cumulative flow function $N(t)$ numbering pedestrians as they pass the cross section and describing the evolution of this number over time (formally, $N(t) = \sum_i 1_{t_i < t}$, with $1_A = 1$ if A holds and $1_A = 0$ otherwise). It is related as the Moskowitz function in the traffic theory (Daganzo 2006). The flow rate is then the derivative in time of the cumulative flow in the continuous limit $J_0 = N'(t)$. On a discrete time interval T , the flow rate is (see Fig. 1)

$$J_T = \frac{\Delta N}{T}, \quad \text{with} \quad \Delta N = N(t) - N(t - T). \quad (2)$$

Another possibility to measure the flow rate of a pedestrian stream is borrowed from fluid dynamics. The flow rate through a facility of width b determined by the average density ρ and the average speed v of a pedestrian stream as

$$J = \rho \ v \ b = J_s b. \quad (3)$$

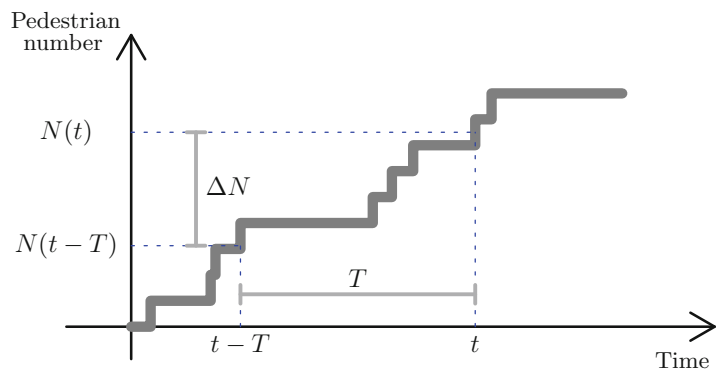
where the *specific flow*

$$J_s = \rho \ v \quad (4)$$

gives the flow per unit width. This relation is also known as *hydrodynamic relation*. In strictly one-dimensional motion, often a line density (dimension: 1/length) is used. Then the *flow* is directly given by $J = \rho v$.

There are several problems concerning the way how velocities, densities, or time gaps are measured and the conformance of the two definitions of the flow rate. The flow according to Eq. (1) is usually measured as a mean value over time at a certain location, while the measurement of the density in Eq. (3) is connected with an instantaneous mean value over space. This can lead to a bias caused by the underestimation of fast moving pedestrians at the average over space compared to the mean value of the flow over time at a single measurement line; see the discussion for vehicular traffic, e.g., in Helbing (2001), Kerner (2004), and Leutzbach (1988). Furthermore most experimental studies measuring the flow rate according to Eq. (3) combine for technical reasons an *average* velocity of a single pedestrian over time with an *instantaneous* density. To ensure a correspondence of the mean values, the average velocity of all pedestrians contributing to the density at a certain instant has to be considered. However this procedure is very time-consuming and not realized in practice up to now. Moreover the fact that the dimension of the test section has usually the same order of magnitude as the extent of the pedestrians can influence the averages over space. These all are possible factors why different measurements can differ in a large way (see discussion in section “Fundamental Diagram”).

Empirical Results of Pedestrian and Evacuation Dynamics,
Fig. 1 Illustration for the measurements of the flow rate on a time interval T with the cumulative flow function $N(t)$



Density

Another way to quantify the pedestrian load of facilities has been proposed by Fruin (1971). The “pedestrian area module” is given by the reciprocal of the density. Thompson and Marchant (1994) introduced the so-called inter-person distance d , which is measured between center coordinates of the assessing and obstructing persons. According to the “pedestrian area module,” Thompson and Marchant call $\sqrt{\frac{1}{\rho}}$ the “average interperson distance” for a pedestrian stream of evenly spaced persons (Thompson and Marchant 1994). An alternative definition is introduced in Helbing et al. (2007) where the local density is obtained by averaging over a circular region of radius R ,

$$\rho(\mathbf{r}, t) = \sum_i f(\mathbf{r}_i(t) - \mathbf{r}), \quad (5)$$

where $\mathbf{r}_i(t)$ are the positions of the pedestrians i in the surrounding of $\mathbf{r}$ and $f(\dots)$ is a Gaussian, distance-dependent weight function. The kernel function f depends on a bandwidth parameter to calibrate. In the Voronoi method (Steffen and Seyfried 2010), the kernels are uniform on the cells given by the Voronoi diagram (see Fig. 2). The Voronoi method is devoid of additional parameter.

In contrast to the density definitions above, Predtechenskii and Milinskii (1971) consider the ratio of the sum of the projection area f_i of the bodies and the total area of the pedestrian stream A , defining the (dimensionless) density $\tilde{\rho}$ as

$$\tilde{\rho} = \frac{\sum_i f_i}{A}, \quad (6)$$

a quantity known as *occupancy* in the context of vehicular traffic. Since the projection area f_i depends strongly on the type of person (e.g., it is much smaller for a child than an adult), the densities for different pedestrian streams consisting of the same number of persons and the same stream area can be quite different.

Besides technical problems due to camera distortions and camera perspective, there are several conceptual problems, like the association of averaged with instantaneous quantities, the necessity to choose an observation area in the same order of magnitude as the extent of a pedestrian together with the definition of the density of objects with nonzero extent, and much more. A detailed analysis on how the way of measurement influences the relations is necessary but still lacking.

Mean Speed

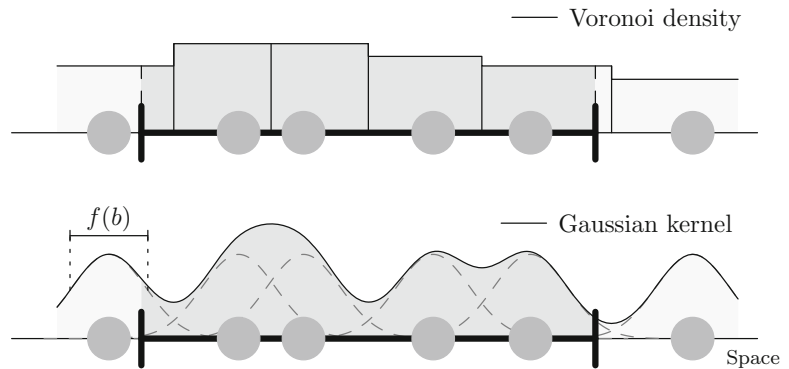
As for flow rate and density, there exist various methods to measure the mean speed. Let us first consider a simple example. A pedestrian goes from a point A to a point B with a given speed v_1 before coming back from B to A with a speed v_2 . We denote L the distance between A and B . What is the mean speed of the pedestrian during the round trip? The time mean speed

$$V = \frac{1}{L + L} (v_1 L + v_2 L) = \frac{1}{2} (v_1 + v_2), \quad (7)$$

is the arithmetic mean of v_1 and v_2 . Yet, the time spent on the way out is $t_1 = L/v_1$ while it is

Empirical Results of Pedestrian and Evacuation Dynamics,

Fig. 2 Examples of Voronoi and Gaussian kernel estimations for the density of pedestrians in one dimension. The Gaussian kernel estimation depends on a bandwidth parameter b to calibrate



$t_2 = L/v_2$ on the return. Therefore the space mean speed is

$$V_H = \frac{1}{t_1 + t_2} (v_1 t_1 + v_2 t_2) = \frac{2}{1/v_1 + 1/v_2}. \quad (8)$$

It is the harmonic mean of v_1 and v_2 . The Jensen inequality shows that the harmonic mean is smaller than the arithmetic mean as soon as v_1 is different from v_2 . For instance, if $v_1 = 1$ and $v_2 = 2$, then $V = 3/2$ while $V_H = 4/3$. Note that only the space mean speed allows to recover the travelled time over the round trip $L/v_1 + L/v_2 = 2 L/V_H$.

Time and space mean speeds do not estimate the same quantity. This statement, closely related to double-loop detector measurement technique, has been early pointed out for traffic flow (Wardrop 1952). However, the difference can be substantial for pedestrian flows as well. It has been empirically estimated up to a factor of 4 (Knoop et al. 2009). The time mean speed is generally measured thanks to crossing times of two consecutive points (as with a double-loop detector). For instance, one can consider that the speeds v_i of pedestrians crossing a section are the length of the section L divided by the difference between the exit t_i^{out} and enter times t_i^{in} into the section (the travel time). Then the arithmetic mean

$$V = \frac{1}{n} \sum_i v_i = \frac{1}{n} \sum_i \frac{L}{t_i^{\text{out}} - t_i^{\text{in}}} \quad (9)$$

corresponds to the time mean speed. It does not allow to recover the mean travel time of the pedestrian into the room, neither the hydrodynamic relation borrowed from fluid dynamics $J = \rho V$. The harmonic mean

$$\begin{aligned} V_H &= \frac{n}{\sum_i 1/v_i} = \frac{n}{\sum_i (t_i^{\text{out}} - t_i^{\text{in}})/L} \\ &= \frac{L}{\frac{1}{n} \sum_i t_i^{\text{out}} - t_i^{\text{in}}} \end{aligned} \quad (10)$$

should be used instead to recover (asymptotically) the space mean speed (Wardrop 1952). Note that the harmonic mean speed corresponds to the

section length divided by the (arithmetic) mean travel time.

The difference between time and space mean speeds can be nicely tackled thanks to trajectories. Let us denote d_i the travelled distance and t_i the travelled time of pedestrian trajectories in the space/time diagram. The speed of pedestrian i is therefore $v_i = d_i/t_i$. A natural way due to Edie (1963) to estimate the mean speed of the trajectories consists in dividing the total travelled distance by the total travel time:

$$V_E = \frac{\sum_i d_i}{\sum_i t_i}. \quad (11)$$

Let us fix now the travelled distance to a fix value d (estimation of the mean speed on a horizontal band in the time/space diagram of the trajectories, see Fig. 3). Such a methodology is the one used above with entry and exit times. One gets the harmonic mean

$$V_E = \frac{\sum_i d}{\sum_i t_i} = \frac{n}{\sum_i 1/v_i}. \quad (12)$$

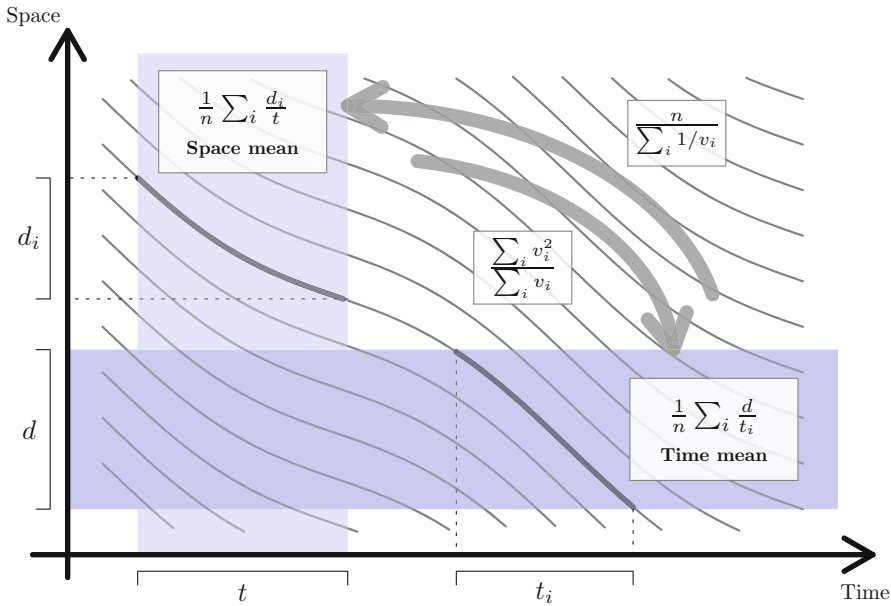
since in this case $t_i = d/v_i$. If we fix the travelled time to a fixed value t (estimation of the speeds thanks to two successive photographs of the system; this corresponds to estimation of the mean speed on a vertical band in the time/space diagram of the trajectories, see Fig. 3), then one gets

$$V_E = \frac{\sum_i d_i}{\sum_i t} = \frac{1}{n} \sum_i v_i, \quad (13)$$

since in this case $d_i = v_i t$. As expected, Edie's definition for the mean speed of trajectories corresponds in any case to the space mean speed.

Collective Effects

One of the reasons why the investigation of pedestrian dynamics has attracted the interest of physicists is the large variety of collective effects and self-organization phenomena that can be



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 3 Illustration for the measurements of the mean speed for trajectories with a fixed time interval (space mean speed, vertical band), a fixed distance (time

mean speed, horizontal band), and the harmonic and contraharmonic means allowing to recover asymptotically one definition from the other

observed. These macroscopic effects reflect the individuals' microscopic interactions and thus give also important information for modelling approaches.

Jamming and Clogging

Jamming and clogging typically occur for high densities. Clogging is typically observed at locations where the inflow exceeds the capacity. This kind of jamming phenomenon does not depend strongly on the microscopic dynamics of the particles. Rather it is a consequence of an exclusion principle: space occupied by one particle is not available for others. Locations with reduced capacity are called *bottlenecks*. Typical examples of bottlenecks are exits (Fig. 4) or narrowings. For practical applications, especially evacuation simulations, a detailed understanding of the conditions under which clogging occurs is important. In Predtechenskii and Milinskii (1978), it is described how this obstruction occurs due to the formation of arches in front of the door under high pressure. This is very similar to the well-known phenomenon of arching occurring in the flow of

granular materials through narrow openings (Wolf and Grassberger 1996).

Other types of jamming occur, e.g., in the case of counterflow where two groups of pedestrians mutually block each other. This also typically happens at high densities and when it is not possible to turn around and move back, e.g., when the flow of people is large.

Density Waves, Stop-and-Go Waves

Density waves in pedestrian crowds can be generally characterized as quasiperiodic density variations in space and time. A typical example is the movement in a densely crowded corridor (e.g., in subway stations close to the density that causes a complete halt of the motion) where phenomena similar to stop-and-go vehicular traffic can be observed, e.g., density fluctuations in longitudinal direction that move backward (opposite to the movement direction of the crowd) through the corridor.

Stop-and-go waves have been observed in real crowds, e.g., on the Jamarat Bridge in Makkah (during the Hajj pilgrimage 2006) (Helbing et al.

Empirical Results of Pedestrian and

Evacuation Dynamics,
Fig. 4 Clogging in front of
a bottleneck



2007), as well as in controlled experiments (Portz and Seyfried 2011; Seyfried et al. 2010a). One surprising difference to stop-and-go traffic observed in vehicular traffic is the separation into standing and slowly moving pedestrians. In contrast, in highway traffic a separation into standing and fast-moving cars is found (neglecting a narrow transition layer). This is usually explained by so-called slow-to-start behavior of cars (Barlovic et al. 1998). Therefore in pedestrian motion, additional mechanisms are at work which have not yet been fully understood.

Lane Formation

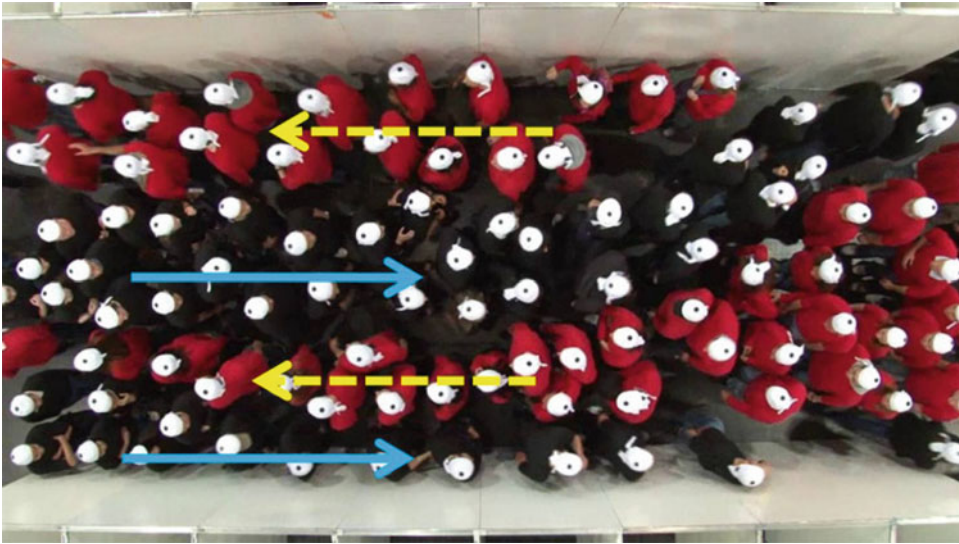
In counterflow, i.e., two groups of people moving in opposite directions, dynamically varying lanes are formed where people move in just one direction (Navin and Wheeler 1969; Oeding 1963; Yamori 1998). In this way, strong interactions with oncoming pedestrians are reduced which is more comfortable and allows higher walking speeds.

The occurrence of lane formation does not require a preference of moving on one side. It also occurs in situations without left or right preference. However, cultural differences for the preferred side have been observed. Although this preference is not essential for the phenomenon itself, it has an influence on the kind of lanes formed and their order.

Several quantities for the quantitative characterization of lane formation have been proposed. Yamori (1998) has introduced a band index which is basically the ratio of pedestrians in lanes to their total number. In Burstedde et al. (2001), a characterization of lane formation through the (transversal) velocity profiles at fixed positions has been proposed. Lane formation has also been predicted to occur in colloidal mixtures driven by an external field (Chakrabarti et al. 2004; Dzubiella et al. 2002; Rex and Löwen 2007). Here an order parameter $\phi = \frac{1}{N} \left\langle \sum_{j=1}^N \phi_j \right\rangle$ has been introduced where $\phi_j = 1$ if the lateral distance to all other particles of the other type is larger than a typical density-dependent length scale and $\phi_j = 0$ otherwise.

The number of lanes can vary considerably with the total width of the flow (see Fig. 5). It is usually not constant and changes in time, even if there are relatively small changes in density. The number of lanes in opposite directions is not always identical. This can be interpreted as a sort of spontaneous symmetry breaking.

Quantitative empirical studies of lane formation are rare (Kretz et al. 2006a; Yamori 1998; Zhang et al. 2012b). At high densities theoretical models predict a jamming transition, i.e., the transition to a state where no movement is possible anymore (gridlock). However, there is no empirical evidence for such a transition, and experiments showed that



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 5 Bidirectional flow with two comparative lanes in each flow direction

the lower density limit for the transition is larger than 3.5 persons/m^2 (Zhang et al. 2012b).

Oscillations

In counterflow at bottlenecks, e.g., doors, one can sometimes observe oscillatory changes of the direction of motion. Once a pedestrian is able to pass the bottleneck, it becomes easier for others to follow in the same direction until somebody is able to pass (e.g., through a fluctuation) the bottleneck in the opposite direction.

Patterns at Intersections

At intersections various collective patterns of motion can be formed. A typical example is short-lived roundabouts which make the motion more efficient. Even if these are connected with small detours, the formation of these patterns can be favorable since they allow for a “smoother” motion.

Emergency Situations, “Panic”

In emergency situations various collective phenomena have been reported that have sometimes misleadingly been attributed to *panic behavior*. However, there is strong evidence that this is not the case. Although a precise accepted definition

of *panic* is missing, usually certain aspects are associated with this concept (Keating 1982). Typically “panic” is assumed to occur in situations where people compete for scarce or dwindling resources (e.g., safe space or access to an exit) which leads to selfish, asocial, or even completely irrational behavior and contagion that affects large groups. A closer investigation of many crowd disasters has revealed that most of the above characteristics have played almost no role and most of the time have not been observed at all (see, e.g., Johnson 1987). Often the reason for these accidents is much simpler, e.g., in several cases the capacity of the facilities was too small for the actual pedestrian traffic, e.g., Luzhniki Stadium Moscow (October 20, 1982), Bergisel (December 4, 1999), and pedestrian bridge Kobe (Akashi) (July 21, 2001) (Tsuji 2003). Therefore the term “panic” should be avoided, *crowd disaster* being a more appropriate characterization. Also it should be kept in mind that in dangerous situations, it is *not* irrational to fight for resources (or your own life), if everybody else does this (Coleman 1990; Mintz 1951). Only from the outside this behavior is perceived as irrational since it might lead to a catastrophe

(Sime 1990). The latter aspect is therefore better described as *nonadaptive behavior*.

We will discuss these issues in more detail in section “Guidelines and Legal Requirements for Evacuation Processes.”

Fundamental Diagram

Overview

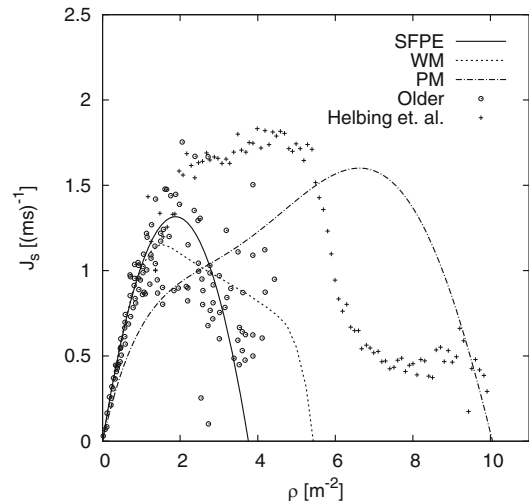
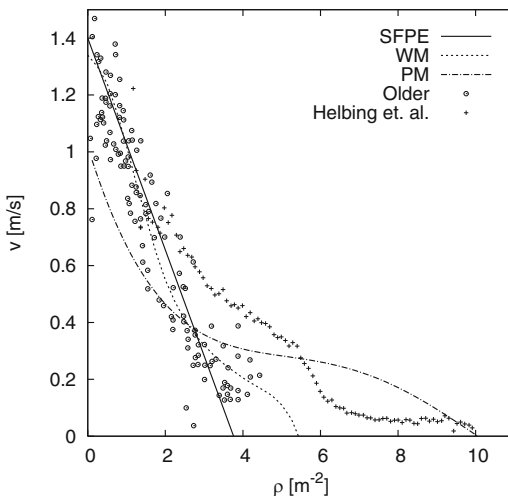
The fundamental diagram describes the relation between density ρ and flow rate J . The name already indicates its importance, and naturally it has been the subject of many investigations. Due to the hydrodynamic relation, there are three equivalent forms, $J_s(\rho)$, $v(\rho)$, and $v(J_s)$, which are basic input in applications for the design and dimensioning of pedestrian facilities (Fruin 1971; Nelson and Mowrer 2002; Predtechenskii and Milinskii 1978). Furthermore it is a quantitative benchmark for models of pedestrian dynamics (Daamen et al. 2002; Kirchner et al. 2004; Meyer-König et al. 2002; Seyfried et al. 2006).

Figure 6 shows various fundamental diagrams used in planning guidelines and measurements of two selected empirical studies representing the overall range of the data. All diagrams agree in

one characteristic: velocity decreases with increasing density. However, the comparison reveals that specifications and measurements disagree considerably. In particular the maximum of the function giving the capacity $J_{s,\max}$ ranges from 1.2 (ms)^{-1} to 1.8 (ms)^{-1} , the density value where the maximum flow rate is reached ρ_c ranges from 1.75 m^{-2} to 7 m^{-2} , and, most notably, the density ρ_0 where the velocity approaches zero due to overcrowding ranges from 3.8 m^{-2} to 10 m^{-2} . Several explanations for these deviations have been suggested, including cultural and population differences (Chattaraj et al. 2009; Helbing et al. 2007; Morrall et al. 1991), differences between uni- and multidirectional flow (Lam et al. 2003; Navin and Wheeler 1969; Pushkarev and Zupan 1975), short-ranged fluctuations (Pushkarev and Zupan 1975), influence of psychological factors given by the incentive of the movement (Predtechenskii and Milinskii 1978), the type of traffic (commuters, shoppers) (Oeding 1963), different types of facilities, etc.

Classical Experiments

In this section, we mainly refer to uni- and bidirectional pedestrian flow on planar facilities like sidewalks, corridors, or halls. Several factors such



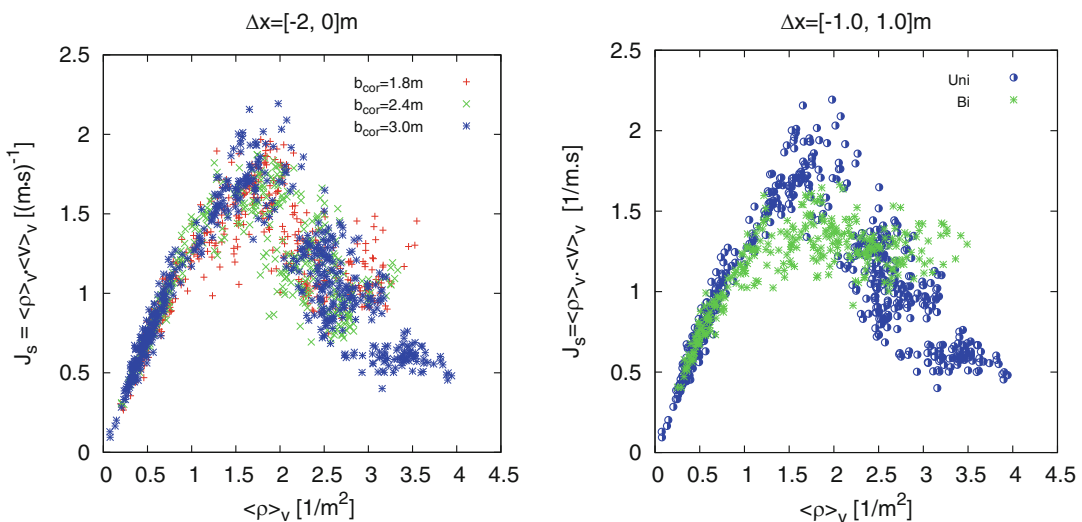
Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 6 Fundamental diagrams for pedestrian movement in planar facilities. The lines refer to specifications according to planning guidelines (SFPE Handbook

(Nelson and Mowrer 2002)), Predtechenskii and Milinskii (PM) (1978), Weidmann (WM) (1993). Data points give the range of experimental measurements (Older 1968; Helbing et al. (2007))

as measurement methods, width of facility, gradients, and pedestrian characteristics have been considered to study the difference of the fundamental diagrams.

For the movement of pedestrians along a line, the speed for walking pedestrians depends linearly on the step size (Weidmann 1993) and the inverse of the density (Seyfried et al. 2005). The internal friction, other lateral interference, and the curvature effects of path have no influence on the fundamental diagram at the density domains considered (Seyfried et al. 2005; Ziemer et al. 2016). However, a different relation is obtained across cultures (Chattaraj et al. 2009; Liu et al. 2009). The speed of Indian is less dependent on density than the speed of German. Instead of changing continuously, the adaptation time takes three discrete values (Jelić et al. 2012) in France. For the group composed of young students with nearly the same age, only two adaptation times are observed (Cao et al. 2016). Especially, the density under ship trim or heeling conditions is not the main factor that affects pedestrian speed. Trim angles show larger impact on speed, and the pedestrian is more likely to be influenced by the front neighbor rather than other pedestrians in front (Sun et al. 2018).

Actually, pedestrians have freedom to move in two dimensions. Corridors are simple and common elements in almost all types of pedestrian facilities designed for uni- and bidirectional pedestrian flow. Different measurement methods mainly influence the range of the fluctuations of the fundamental diagram, which can be unified in one diagram for specific flow for the same type of facility but different widths (Hankin and Wright 1958; Navin and Wheeler 1969; Oeding 1963; Older 1968; Zhang et al. 2011) (see Fig. 7 left). However, Navin and Wheeler observed in narrow sidewalks more orderly movement leading to slightly higher specific flows than for wider sidewalks (Navin and Wheeler 1969). Weidmann (1993) neglected differences between uni- and multidirectional flows in accordance with Fruin, who states in his often cited book (Fruin 1971) that the fundamental diagrams of multidirectional and unidirectional flow differ only slightly. Nevertheless, clear differences in fundamental diagrams of uni- and bidirectional pedestrian streams are observed in Fig. 7 (right) (Lam et al. 2003; Navin and Wheeler 1969; Zhang et al. 2012b). Referring to the influence of flow ratio of bidirectional stream, it seems not an independent factor influencing the fundamental diagram,



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 7 Fundamental diagrams for pedestrian movement in a straight corridor. The left results are obtained from unidirectional movement in corridors with

three different widths (Zhang et al. 2011), while in right the fundamental diagrams of uni- and bidirectional flow are compared (Zhang et al. 2012b)

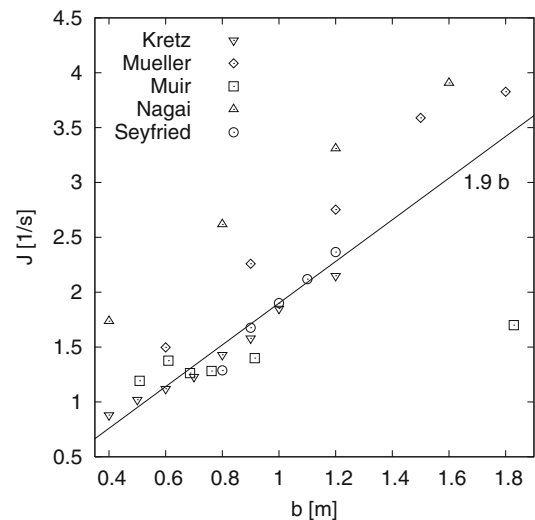
and no consensus is reached up to now (Alhajyaseen et al. 2011; Feliciani and Nishinari 2016; Lam et al. 2002; Transportation Research Board 2000). Surprisingly some studies found that the sum of flow and counterflow in corridors is larger than the unidirectional flow, and for equally distributed loads, it can be twice the unidirectional flow (Kretz et al. 2006a).

Bottleneck Flow

To describe the performance of bottlenecks, two cases have to be distinguished. In the free flow case, the flow through the bottleneck is equivalent to the incoming flow. If the incoming flow exceeds the capacity – the maximal possible flow rate – of the bottleneck, a congestion occurs, and the density in front of the bottleneck increases. This is named congested case. In the congested case, the density in front of the bottleneck is higher than inside the bottleneck (Daamen and Hoogendoorn 2006; Predtechenskii and Milinskii 1978; Seyfried et al. 2007). The measured maximal flow rate at bottlenecks can exceed the maximum of the empirical fundamental diagram. In case of a congestion, pedestrians could cooperate or compete while entering the bottleneck. The competition is triggered by a reward (Mintz 1951) which could be the survival in case of a threat or an uncritical but limited resource like a seat in a carrier. One strategy to successfully enter a bottleneck in a competitive situation is pushing. But pushing could lead to blockages (clogs) limiting the flow through the bottleneck. While the phenomenon of blockages at bottlenecks appears also in systems of inanimate particles (Zuriguel 2014; Zuriguel et al. 2014), the influence of rewards to the degree of competition or to the strength of the pushing is connected with sociopsychological factors (Sieben et al. 2017).

Here we focus on pedestrian movement in cooperative situations. One of the most important practical questions is how the capacity of the bottleneck increases with the width, which is already studied in the beginning of the twentieth century (Dieckmann 1911; Fischer 1933). A stepwise increase of capacity with the width is up to now an assumption of several building codes and design recommendations (MVStättV –

Erläuterungen 2005; Pauls et al. 2006). The empirical results in Hoogendoorn and Daamen (2005) and Hoogendoorn et al. (2006) also imply a stepwise increase of capacity. However, the investigation was restricted to two values of the width. In contrast, the study (Seyfried et al. 2007) considered more values of the bottleneck width in different laboratory experiments (Kretz et al. 2006b; Muir et al. 1996; Müller 1981; Nagai et al. 2006; Seyfried et al. 2007) (see Fig. 8). It is found that the flow does not necessarily depend on the number of lanes. The capacity is given by the maximum of the fundamental diagram. Fruin (1971), Nelson and Mowrer (2002), Predtechenskii and Milinskii (1978), Tian et al. (2012), and Weidmann (1993) assume that the flow is a linear function of the bottleneck width. The exact geometry of the bottleneck is of only minor influence on the flow (Seyfried et al. 2007), while a high initial density in front of the bottleneck can increase the resulting flow values (Nagai et al. 2006). However, it is found later that shorter bottlenecks provide higher flow than longer ones (Liddle et al. 2009; Schadschneider and Seyfried



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 8 Influence of the width of a bottleneck on the flow. Experimental data (Muir et al. 1996; Müller 1981; Nagai et al. 2006; Seyfried et al. 2007) of different types of bottlenecks and initial conditions. All data are taken under laboratory conditions where the test persons are advised to move normally

2011; Seyfried et al. 2010b). The width of the passage (Rupprecht et al. 2011) and the existence of the obstacle (Helbing et al. 2005; Yanagisawa et al. 2009) in front of bottleneck also influence the flow through bottleneck. The total flow rate at bottlenecks with bidirectional movement is higher than it is for unidirectional flows (Helbing et al. 2005). Lanes in bottlenecks are solely a boundary effect, and in case of wide bottlenecks, only two lanes at the boundaries are observable (Liddle et al. 2009). Besides, population with 5% disabled pedestrians leads to a lower flow and that with mainly children leads to the highest flow (Daamen and Hoogendoorn 2010).

Empirical studies of competitive situation where people push while entering the bottleneck are Garcimartín et al. (2016), Helbing et al. (2005), Muir et al. (1996), Müller (1981), and Yanagisawa et al. (2009). In Müller (1981) soldiers entered a short bottleneck with the command “everyone for themselves.” Temporal blockages are reported, but with an increase of the width, their frequency decreases. Muir et al. (1996) studied emergency evacuation in aircrafts. The motivation was varied by a reward, and the dependence of the evacuation time on the width of a bottleneck, here the galley unit, was studied. A comparison of the relation between evacuation time and the width of the bottleneck with and without rewards showed a crossover. While for narrow bottlenecks the evacuation time for high motivation was lower than for normal motivation, the opposite is the case for wider bottlenecks. The high motivation improves the performance and minimizes the evacuation time. These two studies are in conformance showing that as long as no blockages occur, a higher motivation could increase the flow through a bottleneck. Blockages are only observable at narrow bottlenecks which may limit the performance. Experimental results on how a column in front of a narrow bottleneck could improve the flow under high competition can be found in Helbing et al. (2005) and Yanagisawa et al. (2009).

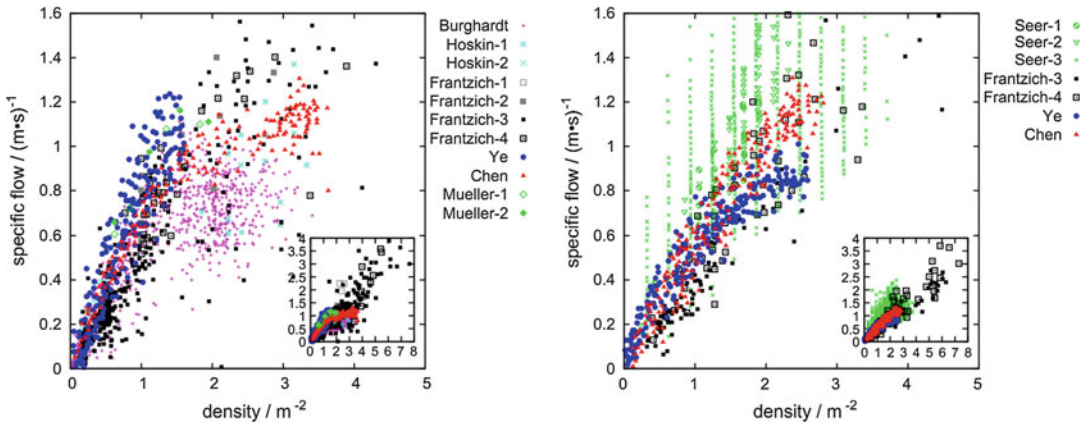
To describe the intermittent flow in case of blockages quantitatively, knowledge about the probability of the occurrence of blockages is necessary. A deeper analysis of the statistical

properties of this phenomenon was studied in Garcimartín et al. (2016). Cumulative distribution functions of the time gaps were analyzed for two door width and three degrees of competition. They showed that the distribution of the time lapses between the passages of two consecutive pedestrians displays heavy-tailed distributions.

Stairs

Stairs are significant elements in most evacuation scenarios like in multistory or high-rise buildings and are a major determinant for the evacuation time. Due to their physical dimension which is often smaller than other parts of a building or due to a reduced walking speed, stairs generally have to be considered as bottlenecks for the flow of evacuees. There are studies on various details, mostly the free speed, of motion on stairs in dependence of the incline (Fruin 1971; Fujiyama and Tyler 2004a; Fujiyama and Tyler 2004b; Graat et al. 1999), conditions (comfortable, normal, dangerous) (Predtetschenski and Milinski 1971), age and sex (Fruin 1971), tread width (Frantzich 1996), the length of a stair (Kretz 2007), and in consideration of various disablement (Boyce et al. 1999).

Compared to the movement on flat terrain, there are more degrees of freedom (e.g., upward and downward movement, the influence of riser height and tread length, exhaustion effects for upward motion, etc.) which influence the fundamental diagram of movement on stairs. In four well-known planning handbooks for pedestrian facilities and evacuation routes (Fruin 1971; Nelson and Mowrer 2002; Predtechenskii and Milinskii 1978; Weidmann 1993), different fundamental diagrams for stairs are presented. Weidmann, Predtechenskii and Milinskii, and Fruin distinguish between up- and downward motion and do not consider the angle of stairs, while Nelson and Mowrer focus on different angles and not on direction. Burghardt et al. (2013) collected fundamental diagram for up- and downward movement on stairs from literatures (Burghardt et al. 2010; Chen et al. 2010; Frantzich 1994, 1996; Graat et al. 1999; Hankin and Wright 1958; Hoskin and Spearpoint 2004; Kretz et al. 2008; Lam and Cheung 2000; Seer



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 9 Fundamental diagram for pedestrian movement on stairs (left, downward; right, upward) (Burghardt et al. 2013)

et al. 2008; Tanaboriboon et al. 1986; Templer 1992; Ye et al. 2008) (see Fig. 9). These empirical data are obtained from different conditions (e.g., different locations and slopes of the stairs, different measurement methods, etc.). However, a linear increase of the specific flow is observable with minor differences of the incline at low densities for both up- and downward motions. The flow rate decreases with increasing slope of stairs (Burghardt et al. 2013; Graat et al. 1999). In the experiment of Burghardt et al. (2013), the density range where the specific flow reaches the maximum is larger for downward than for upward motion. For downward motion the specific flow at the external staircase fades into a plateau for densities higher than 2.0 m^{-2} . For the stairs located in the lower tier of the grandstand, the specific flow increases with higher densities, while in the upper tier, a decrease of flow appears.

Other Geometries

Besides the above mentioned elements, fundamental diagram of pedestrian flow through other structures like T-junction crossing (see Fig. 10) is also studied recently. In T-junctions, bottleneck flow, merging flow, and split flow are all possible to take place in different situations. Around corners, it is still not known how the effective width of the corridor reduces and changes with increasing inflow. The existing empirical data shows that the fundamental diagrams of streams in front and

behind the turning at the corner agree well and are in accordance with that from T-junction flow behind the merging (Zhang et al. 2012a). The fundamental diagram obtained in the mainstream is different from that in branches (Zhang et al. 2013). At crossing pedestrians from different directions have different preferred velocities with different orientations. Thus, the usual way for measuring local velocity or flow rate is not applicable, and adapted definitions have been proposed (Cao et al. 2017; Lian et al. 2015b). Surprisingly, the fundamental diagrams for bidirectional flow and four-directional crossing flow show no apparent difference in the density ranges covered in the experiments. However, it indicates that different measurement methods, motivations of test persons, and experimental setup may lead to large differences.

Evacuations and Crowd Disasters

We have focused in the previous sections on the main variables and empirical results for pedestrian motion in rather simple scenarios. As we have seen, there are many open questions where no consensus has been reached, sometimes even about the qualitative aspects. This becomes even more relevant for full-scale descriptions of evacuations from large buildings, e.g., train stations, malls, theater, or stadium, and/or again for crowd disasters.



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 10 Crossing with four entrances and equal flow from and to all branches

Evacuation

Evacuation is the movement of persons leaving a place due to a danger, threat, or occurrence of a disastrous event. In normal situations an evacuation is called “egress” instead. In a normal situation, visitors of a building might have complex itineraries which are usually represented by origin-destination matrices. In the case of an evacuation, however, the aims and routes are known and usually the same, i.e., the exits and the egress routes. This is the reason why an evacuation process is rather strictly limited in space and time, i.e., its beginning and end are well defined (sound of the alarm, initial position of all persons, safe areas, final position of all persons, and the time, the last person reaches the safe area). Five different phases can be distinguished (Hamacher and Tjandra 2002; MSC-Circ.1033 n.d.; Purser and Bensilium 2001) during an evacuation: (1) detection time, (2) awareness time, (3) decision time, (4) reaction time, and (5) movement time. In International Maritime Organization (IMO) regulations (MSC-Circ.1033 n.d.; MSC-Circ.1166 2005), the first four are grouped together into *response time*. Usually, this time is called *pre-movement time*.

Knowledge about bottleneck capacities (i.e., flows through doors and on stairs) is very

important when assessing the layout of a building with respect to evacuation. The purpose of empirical data in the context of evacuation processes (and modelling in general) is threefold (Galea 2002; ISO-TR-13387-8-1999 1999): (1) identify parameters (factors that influence the evacuation process, e.g., bottleneck widths and capacities); (2) quantify (calibrate) those parameters, e.g., flow through a bottleneck in persons per meter and second; and (3) validate simulation results, e.g., compare real evacuation times to simulation or calculation results. However, real data of evacuation are relatively scarce. Furthermore, real cases of evacuation have generally very specific contexts. Experimental evacuations present also important difficulties. Due to practical, financial, and ethical constraints, an evacuation trial cannot be realistic by its very nature. Therefore, an evacuation exercise does not generally convey the increased stress of a real evacuation.

Guidelines and Legal Requirements for Evacuation Processes

The evacuation of a building can either be an isolated process (due to fire restricted to this building, a bomb threat, etc.), or it can be part of the evacuation of a complete area. In the following we focus on the single building evacuation. For the

evacuation of complete areas, e.g., because of flooding or hurricanes, see (Revi 2006) and references therein.

In many countries there is no strict criterion for the maximum evacuation time of buildings. The requirements in legal building regulations, standards, or guidelines are usually based on specifications of maximal escape path length as well as on maximal number of occupants permitted in combination with minimum exit widths. These specifications can vary from country to country significantly, see (Forell et al. 2010). These variations base on different assumptions with respect to capacity as well as the dependence of the capacity on the width (see section “[Flow Rate](#)”).

Safety during a building evacuation is furthermore assured by the quality of the escape routes or building services. This includes, for example, measures to ensure smoke and fire-free escape routes for a certain time or the planning of doors that can be opened in the direction of the evacuation movement. Building services to enhance the evacuation include automatic fire detection systems, alarming systems, or sprinkler systems.

For passenger ships, a distinction between high-speed craft (HSC), Ro-Ro passenger ferries, and other passenger vessels (cruise ships) is made. High-speed craft do not have cabins, and the seating arrangement is similar to aircraft. Therefore, there is a separate guideline for HSC (MSC-Circ.1166 2005). For an overview over IMO’s requirements and the historical development up to 2001 see (Dogliani 2002). In addition to the five components for the overall evacuation time listed in section “[Evacuation](#),” there are three more specifics for ships: (6) preparation time (for the life-saving appliances, i.e., lifeboats, life rafts, davits, chutes), (7) embarkation time, and (8) launching time. Therefore, the evacuation procedure on ships is more complex than for buildings.

For high-speed craft, the time limit is 17 min for evacuation (Code 2000), for Ro-Ro passenger ships, it is 60 min (MSC-Circ.1033 n.d.), and for all other passenger ships (e.g., cruise ships), it is 60 min if the number of main vertical zones is less or equal than five and 80 min otherwise (MSC-Circ.1033 n.d.).

For aircraft, the approach can be compared to that of HSC. Firstly, an evacuation test is mandatory, and there is a time limit of 90 s that has to be complied to in the test (FAA 1990).

A number of real evacuations have been investigated and reports are publicly available. Among them the most recent ones for buildings are Beverly Hills Club (Bryan 1995), MGM Grand Hotel (Bryan 1995), retail store (Ashe and Shields 1999), department store (Abe 1986), World Trade Center (Grosshandler et al. 2003) and www.wtc.nist.gov, high-rise buildings (Pauls 1971; Seeger and John 1978), and theater (Weckman and Mannikkö 1999). Furthermore there are reports on the high-speed craft “Sleipner” (NMJP 2000) and for trains (Galea 2002; Schneider et al. 2006). For aircraft, overviews can be found in (Owen et al. 1998) and (Muir 2002). Other studies consider, exit width variation (Muir et al. 1996) and double-deck aircraft (Jungermann and Göhlert 2000).

Crowd Disasters

A non-exhaustive list of examples for crowd disasters include the Victoria Hall Disaster (1883) (Predtechenskii and Milinskii 1978), the crowning ceremony of Tsar Nicholas II (1896) (Schelajew et al. 2000), the Iroquois Theater in Chicago (1903), or more recently a governmental Christmas celebration in Aracaju (2001), the distribution of free saris in Uttar Pradesh (2004), the opening of an IKEA store in Jeddah (2004), crowd movements in football stadium (Luzhniki Stadium in 1982, Hillsborough Stadium in 1989, Johannesburg Stadium in 2001, Abidjan Stadium in 2009), the pilgrimage to Mecca (1990, 2006, 2015), or again during festivals (e.g., Spring Festival in Peking in 2004, Love Parade in Duisburg in 2010, Water Festival in Phnom Penh in 2010, Patna Gandhi’s Maidan Festival in 2014).

Although empirical data on crowd disasters exist, e.g., in the form of reports from survivors or even video footage, it is almost impossible to derive quantitative results from them. Qualitatively, most of crowd disasters are connected with overfilled areas (rooms or other structural installations) and to small exits or other to small

parts of a pedestrian facility. Crowd disasters occur in situation with and without external threats putting into question the role of irrational behavior due to fear. First jamming occurs which could develop to strong pushing and shoving. Rewards could increase the possibility that a simple jam evolve to a pushing crowd (Mintz 1951; Muir et al. 1996; Sieben et al. 2017). There could be very high rewards like the reward to survive (e.g., in case of a fire), but also seemingly small rewards (e.g., reaching a certain attraction) could initiate the pushing of individuals in a crowd.

The concept of “panic” and its relevance for crowd disasters are not scientifically proven and discussed rather controversially. Panic is usually used to describe irrational, selfish, and unsocial behavior. In the context of evacuations, empirical evidence shows that this type of behavior is rare (ASA 2002; Clarke 2002; Keating 1982; Sime 1990). On the other hand, there are indications that fear might be “contagious” (de Gelder et al. 2004). Furthermore related concepts like “herding” and “stampede” seem to indicate a certain similarity of the behavior of human crowds with animal behavior (Johnson 1987; Saloma 2006). The origin of this view on panic behavior in crowds could be traced back to LeBon and the riots of the Paris commune (Le Bon 1895). LeBon developed a theory to explain these riots basing on the idea of a certain collectivity. Nowadays social psychologists rate the concept of LeBon and the theory of collective behaviour initiating a panic as an oversimplification with a weak (or better no) scientific background (McPhail and Tucker 2003; Schweingruber and Wohlstein 2005).

This terminology is quite often used in the public media. It seems to be natural that herding, for instance, exists in certain situations, e.g., limited visibility due to failing lights or strong smoke when exits are hard to find. However, notion of herding or stampede is difficult to quantify and measure in human crowds.

Causes for congestions should be discriminated from causes for death. Some police reports note asphyxiation as one cause of death. However, it is not clear how and why the victims suffocated. In cases of heavy pushing, eyewitness reported (e.g., from the Love Parade in Duisburg) that

people were lying on the ground not being able to stand up again, even with the help of other people. Several persons suffocated since the pressure due to the pushing was too high. Why people fell to the ground (being pushed or stumble) is still unclear. Fatalities occurred also at stampedes where people were overrun.

Controlled Experiments

To understand and thereupon to model pedestrian dynamics, reliable empirical data is needed. The dynamics of a person, a group, or crowd has a large variety of influencing aspects. Personal factors are the plan or goal in mind, the local knowledge of the place, experience of the current situation, constitution, motivation, stress level or cultural affiliation, and much more. Some of these factors in turn are influenced by the age, gender, weight, body size, or disabilities (from visual aid to wheelchair). Interpersonal aspects are, for example, the size and density of the crowd, flow directions, atmosphere, dynamic, togetherness, and grouping (family or friends). The personal and the interpersonal factors are influenced by the surrounding like the facility design (e.g., complexity of a building, signage), lighting conditions, or available routes.

Thus motion and behavior are influenced by a variety of factors, and furthermore crowds can be rather inhomogeneous so that for the investigation of single parameters, other influencing aspects should be kept as constant as possible. This can be facilitated by controlled experiments, although the motivation of the participants can decrease over a long-term study, or an adaptation to the procedure of the experiment can happen. But only with laboratory experiments one is able to selectively analyze parameters under well-defined constant conditions without undesired influences, vary these parameters of interest as needed, and adjust them to irregularly or seldom situations in field studies like a specific quantity of disabled people (Christensen et al. 2014) or very high densities (Shi et al. 2015). For self-initiated experiments, the location and the structure of the test persons (e.g., culture, fitness, age, gender, size)

can be determined. Moreover, optimal conditions in artificial environments ensure the extraction of high precise data with low error (Boltes and Seyfried 2013) and enable the measurement of quantities like psychological aspects (Sieben et al. 2017) not possible to survey in observations.

For field studies collecting data is cheaper whereby the precision is lower and extracting meaningful information is much harder. The situation at field studies is more realistic, and some parameters influencing pedestrian dynamics are difficult to set up artificially (e.g., as mentioned in section “[Guidelines and Legal Requirements for Evacuation Processes](#)” the stress of a real evacuation in comparison to an evacuation exercise).

To give the possibility to reproduce findings and to allow further analysis of data extracted from elaborate experiments, sharing this data can push forward the field of pedestrian dynamics. An open access data archive can be found at Forschungszentrum Jülich (2017). This is a good starting point for experimental studies on pedestrian dynamics.

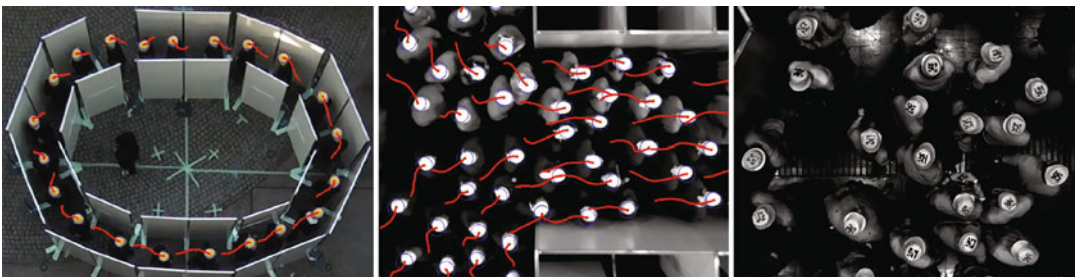
Measurement Techniques

The level of detail of the extracted information from observations and experiments presented in the literature varies. For getting an impression of the overall movement of a crowd, determining abnormal behavior, or separating the crowd in areas of different activities, the optical flow can help (Ali and Shah 2007; Krausz and Bauckhage 2012; Pathan and Richter 2015). Also the

calculation of the velocity in a certain area is possible without detecting single pedestrians. An estimation of the density is feasible as well (Morerio et al. 2012; Ryan et al. 2014). A high level of detail detects and tracks the skeleton of a person. First studies analyzing precise motion sequences already started with the chronophotography in the nineteenth century (Muybridge 1887). Today the Microsoft Kinect for gaming and motion capturing systems in film productions for steering virtual characters are able to do the skeletal detection and tracking automatically in real time (Gmitterko and Liptak 2013).

For the analysis of peoples’ motion, the highest possible level would be the best, but for crowds with high density at present, no system is able to track the full locomotor system of each subject. This is one reason why up to now most of the models simulating pedestrian dynamics do not consider the motion of all parts of a body, although taking, for example, the gait into account may enhance the quality of the simulation results (Chraibi et al. 2010; Seitz and Köster 2012; von Sivers and Köster 2015). Therefore most experimental data provided for analysis and model pedestrian dynamics are trajectories of every single pedestrian, sometimes enriched with additional global (e.g., distribution of age and gender) or individual (e.g., body size, head, or shoulder orientation) information.

By coding each participant individually as shown in the right of Fig. 11, it is possible to identify every person and fuse everyone’s trajectory with static information like human factors, age, or gender gathered by handed-out questionnaires



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 11 Laboratory experiments using colored caps, structured marker with a black dot, or code

marker. For the left and middle picture, the determined position is encircled in blue, and the red path is the way of every person in the last second

(Boltes et al. 2017b; Bukáček et al. 2014; Mehner et al. 2016; Stuart et al. 2013). In addition a unique code of a person during a series of experiments allows the identification of individuals across different runs. But also without individualized trajectories, a questionnaire study is able to reveal how people perceive and evaluate a situations.

To track more than only the head or shoulder of a person in a crowd, the use of inertial sensors started. Inertial measurement units consist of an accelerometer, gyroscope, and magnetometer measuring the acceleration, angular rate, and magnetic field of a moving object. With those readings it is possible to keep track of relative changes of the position of body parts (Boltes et al. 2017b; Schumann and Boltes 2017) so that, e.g., the invisible stepping in a crowd can be analyzed.

The use of eye tracker allows, for example, the investigation of the perception of signs, capturing objects which get the most attention or studying the collision avoidance behavior (Kitazawa and Fujiyama 2010).

The captured dynamic sensor data focus up to now on the physical momentum inside the crowd. For a better exploration of the social and psychological character of the crowd and to combine aspects of behavioral biology (Bode et al. 2015) and social psychology (Sieben et al. 2017) with the natural scientist perspective, additional data like electrodermal activities or pulse is helpful.

As mentioned above currently the common extracted data to study pedestrian dynamics in crowded situations are paths of each person. For the extraction of these trajectories, different techniques have been developed. To perform field studies, methods have to be chosen without intervention and preparation of the observed people. These methods often also have to get along with the restriction having to work in real time (e.g., for privacy protection) and using already installed surveillance cameras with a slanted viewing angle (large observation space but high occlusion level). To detect individual persons in crowds, monocular cameras (Nguyen et al. 2016), stereo cameras (Boltes et al. 2011; van Oosterhout et al. 2015), multi-camera systems (Kiss and Szirányi

2013; Pellicano et al. 2017), near-infrared cameras (Hadi et al. 2013), thermographic cameras (Saito et al. 2008), RGB-D sensors (Corbetta et al. 2014; Jafari et al. 2014), or time-of-flight cameras (Benedek 2014) have been used.

Also other systems exist, but according to Teixeira et al. (2010), the best modality across the board for detection and tracking of people is vision (i.e., cameras and other imagers), and computer vision is far ahead from other instrumented modalities especially with respect to spatial-resolution and precision metrics. Nevertheless Dollar stated in Dollár et al. (2012) that also under favorable conditions with pedestrians at least 80 pixels tall, 20–30% of all pedestrians are missed with at most one false alarm every ten images. Also Nguyen writes in 2015 (Nguyen et al. 2016) that effective detectors can only be constructed for applications where a full upright body is visible or an upper body with less deformation. All camera systems have in common that the detection error increases with increasing density.

Prior information helps to enhance the detection rate. For laboratory experiments one is able to mark the participants; thus most of the researchers collecting data from these experiments use utilities to ease the detection. Figure 11 shows three laboratory experiments with different marker types (colored cap, structured marker, and code marker). The detected position of every pedestrian is encircled in blue, while red paths show the position of the last second. For the code marker, the picture shows only one view of a camera grid to enable the automatic readout of the code. Most studies use colored caps (Chen et al. 2017; Daamen and Hoogendoorn 2010; Hoogendoorn et al. 2003a; Lian et al. 2015a; Liu et al. 2009, 2016; Saadat and Teknomo 2011; Shiwakoti et al. 2015a, b; Tian et al. 2012; Tomoeda et al. 2015; Wong et al. 2010). Others utilize structured marker, e.g., for indicating the head (Boltes et al. 2010; Ezaki et al. 2016) or the shoulder direction (Lemercier et al. 2011), or code additional information (Bukáček et al. 2014; Mehner et al. 2016, 2015; Stuart et al. 2013) for a more precise trajectory (Boltes et al. 2016; Bukáček et al. 2015).

Data Extraction

To extract a real metric position from a raster graphics image, an intrinsic and extrinsic calibration has to be performed. The intrinsic calibration determines camera parameters like focal length and principal point and nonlinear parameters describing the lens distortion. This camera resectioning uses a model that never fits perfectly the real distortion of a lens system, but rectifies the image so that one pixel covers ideally an equal volume in real space. The extrinsic calibration calculates the position and the heading of the camera in real world. Thereby a real position can be calculated for a specified distance to the camera (Brown 1971).

For the detection of a person, most researchers use caps to mark a subject. Colored caps are most common. The center of a head-sized segment with pixel colors in a specified color space volume depicts the position of that person. To detect in every frame, the same point on the head caps with labels can be used for more precise trajectories. These structured markers often code additional information like the height, the line of sight, or an individual code. These features can be extracted, e.g., by isolines of the same brightness with appropriate shape (Boltes et al. 2010).

The tracking of detected positions between successive frames can be done by methods like the Kanade-Lucas-Tomasi feature (Bouguet 1999), Kalman filter (Rameshbabu et al. 2012), mean-shift tracker (Allen et al. 2004), or Camshift tracker (Bradski 1998).

For extracting the path of each person in crowds, the viewing angle should be as perpendicular as possible to have minimal occlusion for small focal lengths. From the people only the head or at most the shoulders are visible and thus can be detected. The trajectories of people tracked in that way describe the path of the head including the swaying and not the center of mass. This data can be used for analyzing the stepping locomotion. For other derived quantities, these data have to be filtered. To determine, e.g., the velocity in the main moving direction, the swaying inside the trajectory has to be smoothed (Steffen and Seyfried 2010), best done by considering the

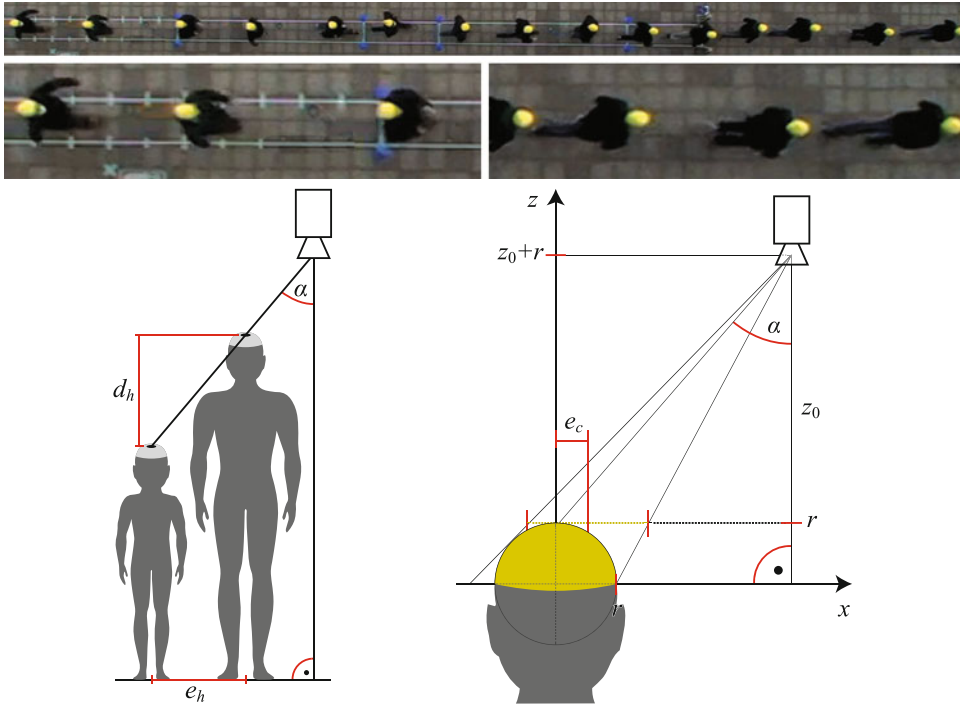
stepping locomotion (Jelić et al. 2012) or the locomotion of the whole body.

To minimize the errors resulting from the perspective view, one should use small angle of view and thus high-mounted cameras to capture the whole experimental area. The small angle of view also decreases the risk of occlusion, and lens systems for large focal lengths generally have a smaller optical distortion error, which are easier to reproduce by a camera model used for undistortion. The accuracy of a detected position is more accurate to the center of an image.

The measurement errors made during the detection of people's position have different reasons (Boltes et al. 2016). Missing (false negative) or surplus (false positive) detections are a big issue for field studies, but are negligible for well-designed laboratory experiments. The camera calibration uses a model that never fits perfectly the real distortion of a lens system so that one has to take this into account especially to the border of images resulting from wide-angle lenses.

As mentioned most researchers use colored caps. Depending on the angle at which these caps are seen by the perspective view of the camera, the resulting geometric center of the color segments belongs to different positions on the head of the persons. For a perpendicular view of the camera, people moving towards the image center are seen from the front, and people moving from the image center to the border are seen from the back as shown on the top of Fig. 12. This leads to a systematic shift of the position to the image center and thus, e.g., to an underestimation of the velocity of a person. This error e_c is sketched in the bottom right of Fig. 12 for an idealized colored cap as hemisphere, a viewing angle α , and a distance to the camera z_0 with an assumed head radius of r . For typical distances z_0 , the error can be estimated by $e_c \approx \alpha \cdot 0.0012 \text{ m}^\circ$. For structured marker this error is not existing.

The perspective view of a camera and its perspective distortion lead to an error of the position, if the distance between the detected head and the camera or the height of the person for plane experiments, respectively, is not known. A pixel in the image may correspond to different heights at



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 12 Top: people with colored caps walking below a camera. The cap on the left is seen from the front and on the right from the back (magnification of the outer part in the second row). Bottom left: the same pixel in

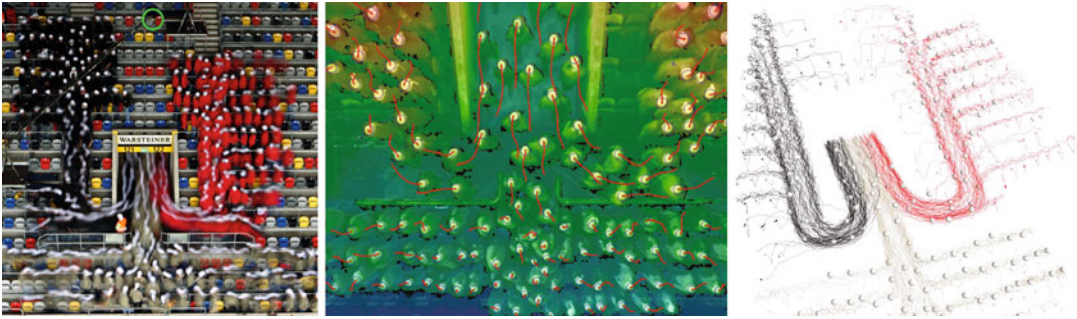
a perspective camera image corresponds to different positions on the ground plane due to differing heights of the persons. Bottom right: the geometric center of the segment of an idealized colored cap as hemisphere in a perspective image is shifted by e_c to the center of the head

different positions on the ground plane as illustrated on the bottom left of Fig. 12. An average height (considering the bobbing) can be chosen, but produces an error e_h of the position on the ground plane of $e_h = |d_h \tan \alpha|$ for a viewing angle α and a height difference d_h . The error variance of the position increases from the image center to the border of the image. To decrease this error, one can use marker coding the height (Boltes et al. 2010) or use techniques like stereo cameras or RGB-D sensors to measure the distance between the camera and the head. These 3D sensors also enable to capture the bobbing of the body movement (Boltes et al. 2011).

Sometimes the area to be observed is too large for one camera, especially if the ceiling height is low so that trajectories of overlapping camera views have to be merged (Boltes et al. 2010). For this purpose the cameras have to be calibrated to the same space and time. For unsynchronized

cameras, a temporal intra-frame shift can decrease the error while fusing the views.

For trajectory extraction on uneven terrains like stairs, knowing peoples' height is no longer sufficient to get a correct position in space. Besides the assumption that all heads are located on a plane in 3D space (Chen et al. 2017), 3D systems like stereo cameras or RGB-D sensors can be used to indicate the 3D position of an image coordinate (Boltes et al. 2011). Figure 13 shows an experiment at a grandstand in a stadium. One stereo camera mounted on a dolly above the port is circled green. The resulting image of one view with an overlaying disparity map and the trajectories of the last second are shown in the middle. The disparity map is color-coded from red to blue according to the distance of 5 m to 11 m to the camera. The right part of Fig. 13 shows all extracted trajectories in 3D space color-coded to the shirt color on the left. White spheres indicate the peoples' position to a fixed time.



Empirical Results of Pedestrian and Evacuation Dynamics, Fig. 13 Left: Experiment with 300 pedestrians leaving a grandstand in a stadium through one port. The green circle points to the position of a stereo camera. Center: Plane view of the stereo recording in front of the

port color-coded to the distance to the camera. The red paths show the head movement in the last second. Right: All trajectories color-coded to the shirt color on the left. White spheres indicate the peoples' position to a fixed time

Conclusions

The increasing importance of safety in public spaces has inspired new approaches to the investigation of pedestrian dynamics. It has been realized that more and better empirical data are needed for a more detailed understanding of the underlying principles. This is also essential for improving modelling approaches and their calibration, e.g., for applications in safety analysis and planning.

Motion and behavior of people and crowds are influenced by a variety of factors (personal and environmental). Due to the difficulties in extracting quantitative data from observations (and potential legal issues) in recent years, laboratory experiments with pedestrians have become the standard tool to collect data for pedestrian streams. They have the advantage of being reproducible and thus allow to study subtle factors like the influence of the cultural background. Besides, controlled experiments allow the investigation of single parameters under well-defined constant conditions.

Currently the most common data gathered at laboratory experiments for studying pedestrian dynamics in crowded situations are paths of each person. For the detection and tracking of these paths, the best modality is vision. Marker allows a robust and precise detection, but lens distortion, perspective view, and peoples' height have to be taken into account. 3D devices allow the extraction

of 3D trajectories. Coding people enables the identification of individuals. Inertial sensors give the possibility to capture the full body motion of pedestrians in crowds, and other devices show the viewing direction or measure the pulse. In short, for further use of the valuable data, the experiments have to be well defined, documented, and publicly available.

Future Directions

The discussion has shown that the problem of crowd dynamics and evacuation processes is far from being well understood. A basic problem that still remains is the empirical basis. Despite the progress in recent years, it is difficult to perform controlled experiments on a sufficiently large scale as in most human systems. These experiments are necessary since data from actual emergency situations is usually not available, at least in sufficient quality. Further progress in the automated and accurate tracking of the motion of individual pedestrians in large crowds without the requirement of equipping the observed people would allow to obtain much more reliable data *in natura*.

Reliable empirical data are also essential for improving theoretical approaches and the validation and calibration of models. Realistic models in combination with detection techniques can be used to improve safety at large-scale events. An

example are evacuation assistants which, in case of emergency, make a prediction for the dynamics of the evacuation process. This allows to identify critical situations (e.g., high-density areas) early on and introduces countermeasures to avoid them. An example for such a system is described in Holl et al. (2014), Holl and Seyfried (2009), and Schadschneider et al. (2013).

Another issue for future research concerns the influence of sociopsychological factors. Although there seems to be a consensus that such factors are relevant and need to be considered in modelling and applications, it is difficult to separate them from the aspects that are related solely to physics, even in laboratory experiments. To make progress in this direction, interdisciplinary research efforts are necessary.

Bibliography

Primary Literature

- Abe K (1986) The science of human panic. Brain Publishing, Tokyo (in Japanese)
- Alhajjaseen WK, Nakamura H, Asano M (2011) Effects of bidirectional pedestrian flow characteristics upon the capacity of signalized cross-walks. *Procedia Soc Behav Sci* 16:526–535
- Ali S, Shah M (2007) A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In: Conference on computer vision and pattern recognition, pp 1–6
- Allen JG, Xu RYD, Jin JS. Object tracking using camshift algorithm and multiple quantized feature spaces. In: Proceedings of the Pan-Sydney area workshop on visual information processing (Darlinghurst, Australia, Australia, 2004), VIP '05, Australian Computer Society, pp 3–7
- Appert-Rolland C, Chevoir F, Gondret P, Lassarre S, Lebacque J-P, Schreckenberg M (2009) Traffic and granular flow '07. Springer, Berlin/Heidelberg
- ASA. In disasters, panic is rare; altruism dominates. Technical report, ASA, Aug 2002
- Ashe B, Shields TJ (1999) Analysis and modelling of the unannounced evacuation of a large retail store. *Fire Mater* 23:333–336
- Barlovic R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793
- Benedek C (2014) 3d people surveillance on range data sequences of a rotating lidar. *Pattern Recogn Lett Special Issue Depth Image Anal* 50:149
- Bode NWF, Holl S, Mehner W, Seyfried A (2015) Disentangling the impact of social groups on response times and movement dynamics in evacuation. *PLoS One* 10:0121227
- Boltes M, Seyfried A (2013) Collecting Pedestrian Trajectories. *Neurocomputing, Special Issue Behav Video* 100:127–133
- Boltes M, Seyfried A, Steffen B, Schadschneider A (2010) Automatic extraction of pedestrian trajectories from video recordings. In: Klingsch WWF, Rogsch C, Schadschneider A, Schreckenberg M (eds) *Pedestrian and evacuation dynamics 2008*. Springer, Berlin/Heidelberg, pp 43–54
- Boltes M, Seyfried A, Steffen B, Schadschneider A (2011) Using stereo recordings to extract pedestrian trajectories automatically in space. In: Peacock RD et al (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 751–754
- Boltes M, Holl S, Tordeux A, Seyfried A, Schadschneider A, Lang U (2017a) Influences of extraction techniques on the quality of measured quantities of pedestrian characteristics. In: Song W, Ma J, Fu L (eds) *Proceedings of pedestrian and evacuation dynamics 2016*. *Collective Dynamics* 1, 0, pp 500–547, 618
- Boltes M, Schumann J, Salden D (2017b) Gathering of data under laboratory conditions for the deep analysis of pedestrian dynamics in crowds. In: 2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)
- Bouguet J-Y (1999) Pyramidal implementation of the Lucas Kanade feature tracker. *OpenCV Documents*
- Boyce KE, Shields TJ, Silcock GWH (1999) Toward the characterization of building occupancies for fire safety engineering: capabilities of disabled people moving horizontally and on an incline. *Fire Technol* 35:51–67
- Bradski GR (1998) Computer vision face tracking for use in a perceptual user interface. *Intel Technol J* 2:1–15
- Brown DC (1971) Close-range camera calibration. *Photogramm Eng* 37:855–866
- Bryan JL (1995) Chapter 3. Behavioral response to fire and smoke. In: DiNenno PJ (ed) *SFPE handbook of fire protection engineering*, 2nd edn. National Fire Protection Association, Quincy, p 263
- Bukáček M, Hrabák P, Krbálek M (2014) Experimental study of phase transition in pedestrian flow. *Transp Res Procedia* 2:105–113
- Bukáček M, Hrabák P, Krbálek M (2015) Experimental analysis of two-dimensional pedestrian flow in front of the bottleneck – experimental analysis of 2d pedestrian flow. In: Chraibi M et al (eds) *Traffic and granular flow '13*. Springer, Heidelberg, pp 93–101
- Burghardt S, Seyfried A, Klingsch W (2010) Improving egress design through measurement and correct interpretation of the fundamental diagram for stairs. In: Panda M, Chattaraj U (eds) *Developments in road transportation*. NIT Rourkela, Odisha, India. Macmillan Publishers India Ltd, pp 181–187
- Burghardt S, Seyfried A, Klingsch W (2013) Performance of stairs – fundamental diagram and topographical measurements. *Transp Res C Emerg Technol* 37:268

- Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A* 295:507–525
- Cao S, Zhang J, Salden D, Ma J, Shi C, Zhang R (2016) Pedestrian dynamics in single-file movement of crowd with different age compositions. *Phys Rev E* 94: 012312
- Cao S, Seyfried A, Zhang J, Holl S, Song W (2017) Fundamental diagrams for multidirectional pedestrian flows. *J Stat Mech Theory Exp* 2017:033404
- Chakrabarti J, Dzubiella J, Löwen H (2004) Reentrance effect in the lane formation of driven colloids. *Phys Rev E* 70:012401
- Chattaraj U, Seyfried A, Chakroborty P (2009) Comparison of pedestrian fundamental diagram across cultures. *Adv Complex Syst* 12(3):393–405
- Chen X, Ye J, Jian N (2010) Relationships and characteristics of pedestrian traffic flow in confined passageways. *Transp Res Rec J Transp Res Board* 2198: 32–40
- Chen J, Lo SM, Ma J (2017) Pedestrian ascent and descent fundamental diagram on stairway. *J Stat Mech Theory Exp* 2017(8):083403
- Chraïbi M, Seyfried A, Schadschneider A (2010) Generalized centrifugal force model for pedestrian dynamics. *Phys Rev E* 82:046111
- Chraïbi M, Boltes M, Schadschneider A, Seyfried A (eds) (2015) *Traffic and granular flow '13*. Springer, Heidelberg
- Chraïbi M, Tordeux A, Schadschneider A, Seyfried A (2018) Modelling of pedestrian and evacuation dynamics. In: *Encyclopedia of complexity and system sciences*, Springer
- Christensen K, Sharifi MS, Stuart D, Chen A, Kim YS, Chen Y (2014) Overview of a large-scale controlled experiment on pedestrian walking behavior involving individual with disabilities. In *The 93rd annual meeting of the transportation research board*
- Clarke LP (2002) Myth or reality? *Contexts* 1:21–26
- Code H. International code of safety for high-speed craft, 2000 (2000 HSC Code). Technical report, International Maritime Organization (IMO), 2000. Resolution MSC97(73)
- Coleman JS (1990) *Foundation of social theory*. Belknap, Cambridge, MA. Chapter 9
- Corbetta A, Bruno L, Muntean A, Toschi F (2014) High statistics measurements of pedestrian dynamics. *Transp Res Procedia* 2:96–104
- Daamen W, Hoogendoorn SP (2006) Flow-density relations for pedestrian traffic. In: Schadschneider A et al (eds) *Traffic and granular flow '05*. Springer, Berlin
- Daamen W, Hoogendoorn SP (2010) Capacity of doors during evacuation conditions. *Procedia Eng* 3(0):53–66. 1st Conference on Evacuation Modeling and Management
- Daamen W, Bovy PHL, Hoogendoorn SP (2002) Modelling pedestrians in transfer stations. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 59–73
- Daganzo CF (2006) On the variational theory of traffic flow: well-posedness, duality and applications. *Netw Heterog Media* 1:601
- de Gelder B, Snyder J, Greve D, Gerard G, Hadjikhani N (2004) Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. *Proc Natl Acad Sci* 101(47): 16701–16706
- Dieckmann D (1911) *Die Feuersicherheit in Theatern*. Jung (München). (in German)
- Dogliani M (2002) An overview of present and under-development IMO's requirements concerning evacuation from ships. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 339–354
- Dollár P, Wojek C, Schiele B, Perona P (2012) Pedestrian detection: An evaluation of the state of the art. *Pattern Anal Mach Intell IEEE Trans* 34(4):743–761
- Dzubiella J, Hoffmann GP, Löwen H (2002) Lane formation in colloidal mixtures driven by an external field. *Phys Rev E* 65:021402
- Edie L (1963) Discussion of traffic stream measurements and definitions. In: Almond J (ed) *Proceedings of the 2nd international symposium theory of traffic flow*, pp 139–154
- El Yacoubi S, Chopard B, Bandini S (eds) (2006) *Cellular automata – 7th international conference on cellular automata for research and industry, ACRI 2006*. Springer, Perpignan
- Ezaki T, Ohtsuka K, Chraïbi M, Boltes M, Yanagisawa D, Seyfried A, Schadschneider A, Nishinari K (2016) Inflow process of pedestrians to a confined space. *Collective Dyn* 1:1–18
- FAA (1990) F. A. A. Emergency evacuation – cfr sec. 25.803. Regulation CFR Sec. 25.803, Federal Aviation Administration
- Feliciani C, Nishinari K (2016) Empirical analysis of the lane formation process in bidirectional pedestrian flow. *Phys Rev E* 94:032304
- Fischer H (1933) *Über die Leistungsfähigkeit von Türen, Gängen und Treppen bei ruhigem, dichtem Verkehr*. Dissertation, Technische Hochschule Dresden. in German
- Forell B, Seidenspinner R, Hosser D (2010) Quantitative comparison of international design standards of escape routes in assembly buildings. In: Klingsch W et al (eds) *Pedestrian and evacuation dynamics 2008*. Springer, Berlin/Heidelberg, pp 791–801
- Forschungszentrum Jülich, Jülich Supercomputing Centre. Data archive of experiments on pedestrian dynamics. <http://ped.fz-juelich.de/dataarchive> 18 Nov 2018
- Frantzich H (1994) A model for performance-based design of escape routes. Technical report 1011, Department of Fire Safety Engineering, Lund Institute of Technology
- Frantzich H (1996) Study of movement on stairs during evacuation using video analysing techniques. Technical report, Department of Fire Safety Engineering, Lund Institute of Technology

- Fruin JJ (1971) Pedestrian planning and design. Elevator World, New York
- Fujiyama T, Tyler N (2004a) An explicit study on walking speeds of pedestrians on stairs. In: 10th international conference on mobility and transport for elderly and disabled people
- Fujiyama T, Tyler N (2004b) Pedestrian speeds on stairs: an initial step for a simulation model. In: Proceedings of 36th Universities' Transport Studies Group Conference
- Galea ER (2002) Simulating evacuation and circulation in planes, trains, buildings and ships using the EXODUS software. In: Schreckenberg M, Sharma SD (eds) Pedestrian and evacuation dynamics. Springer, Berlin/Heidelberg, pp 203–226
- Galea ER (ed) (2003) Pedestrian and evacuation dynamics 2003. CMS Press, London
- Garcimartín A, Parisi DR, Pastor JM, Martín-Gómez C, Zuriguel I (2016) Flow of pedestrians through narrow doors with different competitiveness. J Stat Mech Theory Exp 2016:043402
- Gminterko A, Liptak T (2013) Motion capture of human for interaction with service robot. Am J Mech Eng 1:212–216
- Graat E, Midden C, Bockholts P (1999) Complex evacuation; effects of motivation level and slope of stairs on emergency egress time in a sports stadium. Saf Sci 31:127–141
- Grosshandler W, Sunder S, Snell J (2003) Building and fire safety investigation of the world trade center disaster. In: Galea ER (ed) Pedestrian and evacuation dynamics 2003. CMS Press, London, pp 279–281
- Hadi HS, Rosbi M, Sheikh UU (2013) A review of infrared spectrum in human detection for surveillance systems. Int J Interact Digit Media 1(3):13–20
- Hamacher HW, Tjandra SA (2002) Mathematical modelling of evacuation problems – a state of the art. In: Schreckenberg M, Sharma SD (eds) Pedestrian and evacuation dynamics. Springer, Berlin/Heidelberg, pp 227–266
- Hankin BD, Wright RA (1958) Passenger flow in subways. Oper Res Q 9(2):81–88
- Helbing D (2001) Traffic and related self-driven many-particle systems. Rev Mod Phys 73:1067
- Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: experiments, simulations, and design solutions. Transp Sci 39:1–24
- Helbing D, Johansson A, Al-Abideen HZ (2007) Dynamics of crowd disasters: an empirical study. Phys Rev E 75:046109
- Holl S, Seyfried A (2009) Hermes – an evacuation assistant for mass events. inSiDe 7:60
- Holl S, Schadschneider A, Seyfried A (2014) Hermes: an evacuation assistant for large arenas. In: Weidmann U et al (eds) Pedestrian and evacuation dynamics 2012 (Zürich, 2014). Springer, Berlin/Heidelberg, p 345
- Hoogendoorn SP, Daamen W (2005) Pedestrian behavior at bottlenecks. Transp Sci 39(2):0147–0159
- Hoogendoorn S, Daamen W, Bovy P (2003a) Extracting microscopic pedestrian characteristics from video data. In: TRB2003 annual meeting
- Hoogendoorn SP, Daamen W, Bovy PHL (2003b) Microscopic pedestrian traffic data collection and analysis by walking experiments: behaviour at bottlenecks. In: Galea ER (ed) Pedestrian and evacuation dynamics 2003. CMS Press, London, pp 89–100
- Hoskin KJ, Spearpoint M (2004) Crowd characteristics and egress at stadia. In: Shields TJ (ed) Human Behaviour in Fire. Interscience, London, pp 367–376
- ISO-TR-13387-8-1999 (1999) Fire safety engineering – part 8: life safety – occupant behaviour, location and condition. Technical report, International Organization for Standardization. www.iso.org
- Jafari OH, Mitzel D, Leibe B (2014) Real-time rgb-d based people detection and tracking for mobile robots and head-worn cameras. In: IEEE international conference on robotics and automation (ICRA)
- Jelić A, Appert-Rolland C, Lemercier S, Pettré J (2012) Properties of pedestrians walking in line: fundamental diagrams. Phys Rev E 85:9
- Johnson NR (1987) Panic at “The Who Concert Stampede”: an empirical assessment. Soc Probl 34:362–373
- Jungermann H, Göhlert C (2000) Emergency evacuation from double-deck aircraft. In: Cottam M, Harvey D, Pape R, Tait J (eds) Foresight and precaution. Proceedings of ESREL 2000, SARS and SRA Europe annual conference. A.A. Balkema, Rotterdam, pp 989–992
- Keating JP (1982) The myth of panic. Fire J:57–62
- Kerner BS (2004) The physics of traffic. Springer, Berlin
- Kerner B (2017) Breakdown in traffic networks – fundamentals of transportation science. Springer, Berlin/Heidelberg
- Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization effects and the influence of walking speed in cellular automata models for pedestrian dynamics. J Stat Mech 2004(10):P10011
- Kiss Á, Szirányi T (2013) Localizing people in multi-view environment using height map reconstruction in real-time. Pattern Recogn Lett 34(16):2135–2143
- Kitazawa K, Fujiyama T (2010) Pedestrian vision and collision avoidance behavior: investigation of the information process space of pedestrians using an eye tracker. In: Klingsch W et al (eds) Pedestrian and evacuation dynamics 2008. Springer, Berlin/Heidelberg, pp 95–108
- Klingsch W, Rogsch C, Schadschneider A, Schreckenberg M (eds) (2010) Pedestrian and evacuation dynamics 2008. Springer, Berlin/Heidelberg
- Knoop VL, Daamen W (eds) (2016) Traffic and granular flow '15. Springer, Berlin/Heidelberg
- Knoop V, Hoogendoorn S, van Zuylen H (2009) Empirical differences between time mean speed and space mean speed. In: Appert-Rolland C, Chevoir F, Gondret P, Lassarre S, Lebacque J-P, Schreckenberg M (eds) Traffic and Granular Flow '07. Springer, Berlin/Heidelberg

- Kozlov V, Buslaev A, Bugaev A, Yashina M, Schadschneider A, Schreckenberg M (eds) (2013) *Traffic and granular flow '11*. Springer, Heidelberg
- Krausz B, Bauckhage C (2012) Loveparade 2010: automatic video analysis of a crowd disaster. *Comput Vis Image Underst* 116(3):307–319 Special issue on Semantic Understanding of Human Behaviors in Image Sequences
- Kretz T (2007) *Pedestrian traffic – simulation and experiments*. PhD thesis, Universität Duisburg-Essen, Fachbereich Physik
- Kretz T, Grünebohm A, Kaufman M, Mazur F, Schreckenberg M (2006a) Experimental study of pedestrian counterflow in a corridor. *J Stat Mech* 2006:P10001
- Kretz T, Grünebohm A, Schreckenberg M (2006b) Experimental study of pedestrian flow through a bottleneck. *J Stat Mech* 2006:P10014
- Kretz T, Grünebohm A, Kessel A, Klüpfel H, Meyer-König T, Schreckenberg M (2008) Upstairs walking speed distributions on a long stairway. *Saf Sci* 46(1):72–78
- Lam WHK, Cheung CY (2000) Pedestrian speed/flow relationships for walking facilities in hong kong. *J Transp Eng* 126:343–349
- Lam WHK, Lee JYS, Cheung CY (2002) A study of the bidirectional pedestrian flow characteristics at Hong Kong signalized crosswalk facilities. *Transportation* 29:169–192
- Lam WHK, Lee JYS, Chan KS, Goh PK (2003) A generalised function for modeling bi-directional flow effects on indoor walkways in Hong Kong. *Transp Res A Policy Pract* 37:789–810
- Le Bon G (1895) *The crowd: a study of the popular mind* (Psychologie des Foules). Sparkling Books
- Lemerrier S, Moreau M, Moussaïd M, Theraulaz G, Donikian S, Pettré J (2011) Reconstructing motion capture data for human crowd study. *Lect Notes Comput Sci* 7060:365–376
- Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
- Lian L, Mai X, Song W, Kit YK (2015a) An experimental study on four-directional intersecting pedestrian flows. *J Stat Mech* 2015:P08024
- Lian L, Mai X, Song W, Richard KYK, Wei X, Ma J (2015b) An experimental study on four-directional intersecting pedestrian flows. *J Stat Mech Theory Exp* 2015(8):P08024
- Liddle J, Seyfried A, Klingsch W, Rupprecht T, Schadschneider A, Winkens A (2009) An experimental study of pedestrian congestions: Influence of bottleneck width and length. available from <https://arxiv.org/abs/0911.4350>
- Liu X, Song W, Zhang J (2009) Extraction and quantitative analysis of microscopic evacuation characteristics based on digital image processing. *Physica A* 388(13):2717–2726
- Liu X, Song W, Fu L, Fang Z (2016) Experimental study of pedestrian inflow in a room with a separate entrance and exit. *Physica A* 442:224–238
- McPhail C, Tucker C (2003) Collective behaviour. In: Reynolds L, Herman-Kinney NJ (eds) *Handbook of symbolic interactionism*. Altamira, Walnut Creek, pp 721–741
- Mehner W, Boltes M, Seyfried A (2016) Methodology for generating individualized trajectories from experiments. In: Knoop VL, Daamen W (eds) *Traffic and granular flow '15*. Springer, Berlin/Heidelberg, pp 3–10
- Mehner W, Boltes M, Mathias M, Leibe B (2015) Robust marker-based tracking for measuring crowd dynamics. Springer International Publishing, Cham, pp 445–455
- Meyer-König T, Klüpfel H, Schreckenberg M (2002) Assessment and analysis of evacuation processes on passenger ships by microscopic simulation. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 297–302
- Mintz A (1951) Non-adaptive group behaviour. *J Abnorm Soc Psychol* 46:150–159
- Morerio P, Marcenaro L, Regazzoni CS (2012) People count estimation in small crowds. In: *Advanced video and signal-based surveillance (AVSS)*, 2012 I.E. ninth international conference on, pp 476–480
- Morrall JF, Ratnayake LL, Seneviratne PN (1991) Comparison of central business district pedestrian characteristics in Canada and Sri Lanka. *Transp Res Rec* 1294:57
- MSC-Circ.1033. Interim guidelines for evacuation analyses for new and existing passenger ships. Technical report, International Maritime Organization, Marine Safety Committee, London, June, 6th 2002. MSC/Circ. 1033
- MSC-Circ.1166. Guidelines for a simplified evacuation analysis for high-speed passenger craft. Technical report, International Maritime Organisation 2005
- Muir HC (2002) Airplane of the 21st century: challenges in safety and survivability. In: *Airplane survivability issues in the 21st century*
- Muir HC, Bottomley DM, Marrison C (1996) Effects of motivation and cabin configuration on emergency aircraft evacuation behavior and rates of Egress. *Int J Aviat Psychol* 6(1):57–77
- Müller K (1981) *Zur Gestaltung und Bemessung von Fluchtwegen für die Evakuierung von Personen aus Bauwerken auf der Grundlage von Modellversuchen*. Dissertation, Technische Hochschule Magdeburg
- Muybridge E (1887) *Animal locomotion*, plate 519. Da Capo Press, New York
- MVStättV – Erläuterungen: Musterverordnung über den Bau und Betrieb von Versammlungsstätten, Erläuterungen, Juni 2005. www.is-argebau.de
- Nagai R, Fukamachi M, Nagatani T (2006) Evacuation of crawlers and walkers from corridor through an exit. *Physica A* 367:449–460
- Navin FD, Wheeler RJ (1969) Pedestrian flow characteristics. *Traffic Eng* 39:30–36
- Nelson HE, Mowrer FW (2002) Chapter 14. Emergency movement. In: DiNenno PJ (ed) *SFPE handbook of fire*

- protection engineering. National Fire Protection Association, Quincy, pp 367–380
- Nguyen DT, Li W, Ogunbona PO (2016) Human detection from images and videos: A survey. *Pattern Recogn* 51:148–175
- NMJP. The high-speed craft MS Sleipner Disaster 26 November 1999. Official Norwegian Reports 2000:31, Norwegian Ministry of Justice and Police, Oslo, 2000
- Oeding D. Verkehrsbelastung und Dimensionierung von Gehwegen und anderen Anlagen des Fußgängerverkehrs. Forschungsbericht 22, Technische Hochschule Braunschweig, 1963
- Older SJ (1968) Movement of pedestrians on footways in shopping streets. *Traffic Eng Control* 10:160–163
- Owen M, Galea ER, Lawrence PJ, Filippidis L (1998) AASK – aircraft accident statistics and knowledge: a database of human experience in evacuation, derived from aviation accident reports. *Aeronaut J* 102:353–363
- Pathan SS, Richter K (2015) Pedestrian behavior analysis with image-based method in crowds. In: Chraibi M et al (eds) *Traffic and granular flow '13*. Springer, Heidelberg, pp 187–194
- Pauls JL Evacuation drill held in the b. c. hydro building 26 June 1969. Building Research Note 80, NRCC, September 1971
- Pauls JL, Fruin JJ, Zupan JM (2006) Minimum stair width for evacuation, overtaking movement and counterflow – technical bases and suggestions for the past, present and future. In: Waldau N et al (eds) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin, pp 57–69
- Peacock RD, Kuligowski ED, Averill JD (eds) (2011) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg
- Pellicano N, Aldea E, Hegarat-Masclé SL (2017) Geometry-based multiple camera head detection in dense crowds. In: *Proceedings of 28th British Machine Vision Conference (BMVC) – 5th activity monitoring by multiple distributed sensing workshop*
- Portz A, Seyfried A (2011) Analyzing stop-and-go waves by experiment and modeling. In: Peacock RD et al (eds) *Pedestrian and evacuation dynamics*. Springer Berlin/Heidelberg, pp. 577–586
- Predtechenskii VM, Milinskii AI (1978) Planning for foot traffic flow in buildings. Amerind Publishing, New Delhi. Translation of: Proektirovanie Zhdanii s Uchetom Organizatsii Dvizheniya Lyudskikh Potokov, Stroiizdat Publishers, Moscow 1969
- Predtetschenski W, Milinski A (1971) Personenströme in Gebäuden – Berechnungsmethoden für die Modellierung. Müller, Köln-Braunsfeld
- Purser DA, Bensilium M (2001) Quantification of behaviour for engineering design standards and escape time calculations. *Saf Sci* 38(2):158–182
- Pushkarev B, Zupan JM (1975) Capacity of walkways. *Transp Res Rec* 538:1–15
- Rameshbabu K, Swarnadurga J, Archana G, Menaka K (2012) Target tracking system using kalman filter. *Int J Adv Eng Res Stud* 2:90–94
- Revi A (2006) Pre and post-cyclone & storm surge evacuation & emergency response in India. In: Waldau N et al (eds) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin
- Rex M, Löwen H (2007) Lane formation in oppositely charged colloids driven by an electric field: chaining and two-dimensional crystallization. *Phys Rev E* 75:051402
- Rupprecht T, Klingsch W, Seyfried A (2011) Influence of geometry parameters on pedestrian flow through bottleneck. In: Peacock RD et al (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 71–80
- Ryan D, Denman S, Sridharan S, Fookes C (2014) An evaluation of crowd counting methods, features and regression models. *Comput Vis Image Underst* 130:1–17
- Saadat S, Teknomo K (2011) Automation of pedestrian tracking in a crowded situation. In: Peacock RD et al (eds) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg, pp 231–239
- Saito H, Hagihara T, Hatanaka K, Sawai T (2008) Development of pedestrian detection system using far-infrared ray camera. *SEI Techn Rev* 66:112–117
- Saloma C (2006) Herding in real escape panic. In: Waldau N et al (eds) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin
- Schadschneider A, Seyfried A (2011) Empirical results for pedestrian dynamics and their implications for modeling. *Netw Heterog Media* 6:545–560
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf D (eds) (2006) *Traffic and granular flow '05*. Springer, Berlin
- Schadschneider A, Eilhardt C, Nowak S, Wagoum AK, Seyfried A (2013) Hermes – an evacuation assistant for large sports arenas based on microscopic simulations of pedestrian dynamics. In: Kozlov V et al (eds) *Traffic and granular flow '11*. Springer, Heidelberg, p 287
- Schelajew J, Schelajewa E, Semjonow N (2000) Nikolaus II. Der letzte russische Zar. Bechtermünz, Augsburg
- Schneider U, Kath K, Oswald M, Kirchberger H (2006) Evakuierung und Verhalten von Personen im Brandfall unter spezieller Berücksichtigung von schienengebundenen Fahrzeugen. Technical report 12, TU Wien
- Schreckenberg M, Sharma SD (eds) (2002) *Pedestrian and evacuation dynamics*. Springer, Berlin/Heidelberg
- Schreckenberg M, Wolf DE (eds) (1998) *Traffic and granular flow '97*. Springer, Singapore
- Schumann J, Boltes M (2017) Tracking of wheelchair users in dense crowds. In: 2017 international conference on indoor positioning and indoor navigation (IPIN)
- Schweingruber D, Wohlstein RT (2005) The madding crowd goes to school: myths about crowds in introductory sociology textbooks. *Teach Sociol* 33(2):136–153

- Seeger PG, John R (1978) Untersuchung der Räumungsabläufe in Gebäuden als Grundlage für die Ausbildung von Rettungswegen, Teil III: Reale Räumungsversuche. Technical report T395, Forschungsstelle für Brandschutztechnik an der Universität Karlsruhe (TH)
- Seer S, Bauer D, Brändle N, Ray M (2008) Estimating pedestrian movement characteristics for crowd control at public transport facilities. In: 11th international IEEE conference on intelligent transport systems
- Seitz MJ, Köster G (2012) Natural discretization of pedestrian movement in continuous space. *Phys Rev E* 86:046108
- Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. *J Stat Mech* 2005:P10002
- Seyfried A, Steffen B, Lippert T (2006) Basics of modelling the pedestrian flow. *Physica A* 368:232–238
- Seyfried A, Rupprecht T, Passon O, Steffen B, Klingsch W, Boltes M (2007) Capacity estimation for emergency exits and bottlenecks. In: *Interflam 2007 – conference proceedings*
- Seyfried A, Portz A, Schadschneider A (2010a) Phase coexistence in congested states of pedestrian dynamics. *Lect Notes Comp Sci* 6350:496
- Seyfried A, Boltes M, Kähler J, Klingsch W, Portz A, Rupprecht T, Schadschneider A, Steffen B, Winkens A (2010b) Enhanced empirical data for the fundamental diagram and the flow through bottlenecks. In: Klingsch W et al (eds) *Pedestrian and evacuation dynamics 2008*. Springer, Berlin/Heidelberg, pp 145–156
- Shi X, Ye Z, Shiwakoti N, Li Z (2015) A review of experimental studies on complex pedestrian movement behaviors. In: *CICTP 2015*, pp 1081–1096
- Shiwakoti N, Gong Y, Shi X, Ye Z (2015a) Examining influence of merging architectural features on pedestrian crowd movement. *Saf Sci* 75:15–22
- Shiwakoti N, Shi X, Zhirui Y, Wang W (2015b) Empirical study on pedestrian crowd behaviour in right angled junction. In: *37th Australasian Transport Research Forum (ATRF)*
- Sieben A, Schumann J, Seyfried A (2017) Collective phenomena in crowds – where pedestrian dynamics need social psychology. *PLoS One* 12:1–19
- Sime JD (1990) Chapter. 5. The concept of panic. In: Canter D (ed) *Fires and human behaviour*, vol 1. Wiley, London, pp 63–81
- Song W, Ma J, Fu L (2017) Proceedings of pedestrian and evacuation dynamics 2016. *Collective Dyn* 1(0):618
- Steffen B, Seyfried A (2010) Methods for measuring pedestrian density, flow, speed and direction with minimal scatter. *Physica A* 389:1902–1910
- Stuart D, Christensen K, Chen A, Kim YS, Chen Y (2013) Utilizing augmented reality technology for crowd pedestrian analysis involving individuals with disabilities. In: *ASME 2013 international design engineering technical conferences and computers and information in engineering conference*
- Sun J, Lu S, Lo S, Ma J, Xie Q (2018) Moving characteristics of single file passengers considering the effect of ship trim and heeling. *Physica A* 490:476
- Tanaboriboon Y, Hwa SS, Chor CH (1986) Pedestrian characteristics study in singapore. *J Transp Eng* 112:229–235
- Teixeira T, Dublon G, Savvides A (2010) A survey of human-sensing: methods for detecting presence, count, location, track, and identity. *ACM Comput Surv* 5:1
- Templer JA (1992) *The staircase: studies of hazards, falls, and safer design*. The MIT Press, Cambridge, MA
- Thompson PA, Marchant EW Simulex; developing new computer modelling techniques for evaluation. In: Kashiwagi T (ed) *Fire safety science – proceedings of the fourth international symposium (Interscience Communications Ltd, West Yard House, Guildford Grove, London, 1994)*. The International Association for Fire Safety Science, pp 613–624. ISBN:1-88627-900-4
- Tian W, Song W, Ma J, Fang Z, Seyfried A, Liddle J (2012) Experimental study of pedestrian behaviors in a corridor based on digital image processing. *Fire Saf J* 47:8–15
- Tomoea A, Yanagisawa D, Nishinari K (2015) Escape velocity of the leader in a queue of pedestrians. In: *Traffic and granular flow 2013*. Springer, Berlin/Heidelberg, pp 213–218
- Transportation Research Board (2000) *Highway capacity manual*. Technical report, Transportation Research Board, Washington, DC
- Tsuji Y (2003) Numerical simulation of pedestrian flow at high densities. In: Galea ER (ed) *Pedestrian and evacuation dynamics 2003*. CMS Press, London, p 27
- van Oosterhout T, Englebienne G, Kröse B (2015) RARE: people detection in crowded passages by range image reconstruction. *Mach Vis Appl*, 26(5):561–573
- von Sivers I, Köster G (2015) Dynamic stride length adaptation according to utility and personal space. *Transp Res B Methodol* 74:104–117
- Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2006) *Pedestrian and evacuation dynamics 2005*. Springer, Berlin
- Wardrop J (1952) Some theoretical aspects of road traffic research. *Proc Inst Civ Eng* 1:325–362
- Weckman LS, Mannikkö S (1999) Evacuation of a theatre: exercise vs calculations. *Fire Mater* 23:357–361
- Weidmann U (1993) *Transporttechnik der Fussgänger*. Technical report. Schriftenreihe des IVT Nr. 90, Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau, ETH Zürich
- Weidmann U, Kirsch U, Schreckenberg M (2014) (eds) *Pedestrian and evacuation dynamics 2012 (Zürich, 2014)*. Springer, Berlin/Heidelberg
- Wolf D, Grassberger P (eds) (1996) *Friction, arching, contact dynamics*. World Scientific, Singapore
- Wong SC, Leung WL, Chan SH, Lam WHK, Yung NHC, Liu CY, Zhang P (2010) Bidirectional pedestrian

- stream model with oblique intersecting angle. *J Transp Eng* 136(3):234–242
- Yamori K (1998) Going with the flow: micro-macro dynamics in the macrobehavioral patterns of pedestrian crowds. *Psychol Rev* 105:530–557
- Yanagisawa D, Kimura A, Tomoeda A, Nishi R, Suma Y, Ohtsuka K, Nishinari K (2009) Introduction of frictional and turning function for pedestrian outflow with an obstacle. *Phys Rev E* 80:036110
- Ye J, Chen X, Yang C, Wu J (2008) Walking behavior and pedestrian flow characteristics for different types of walking facilities. *Transp Res Rec J Transp Res Board* 2048:43–51
- Zhang J, Klingsch W, Schadschneider A, Seyfried A (2011) Transitions in pedestrian fundamental diagrams of straight corridors and T-junctions. *J Stat Mech Theory Exp* 2011:06004
- Zhang J, Klingsch W, Rupperecht T, Schadschneider A, Seyfried A (2012a) Empirical study of turning and merging of pedestrian streams in T-junction. In: Fourth international symposium on agent-based modeling and simulation (ABModSim-4)
- Zhang J, Klingsch W, Schadschneider A, Seyfried A (2012b) Ordering in bidirectional pedestrian flows and its influence on the fundamental diagram. *J Stat Mech Theory Exp* 2012:P02002
- Zhang J, Klingsch W, Schadschneider A, Seyfried A (2013) Experimental study of pedestrian flow through a T-junction. In: Kozlov V et al (eds) *Traffic and granular flow '11*. Springer, Heidelberg
- Ziemer V, Seyfried A, Schadschneider A (2016) Congestion dynamics in pedestrian single-file motion. In: *Traffic and granular flow 2015*
- Zuriguel I (2014) Invited review: clogging of granular materials in bottlenecks. *Pap Phys* 6:060014
- Zuriguel I, Parisi DR, Hidalgo RC, Lozano C, Janda A, Gago PA, Peralta JP, Ferrer LM, Pugnaroni LA, Clément E, Maza D, Pagonabarraga I, Garcimartín A (2014) Clogging transition of many-particle systems flowing through bottlenecks. *Sci Rep* 4:7324
- ## Books and Reviews
- Ali S, Nishino K, Manocha D, Shah M (eds) (2013) *Modeling, simulation and visual analysis of crowds: a multidisciplinary perspective*. Springer, New York
- Collective dynamics – a multidisciplinary journal for pedestrian dynamics, vehicular traffic and other systems of self-driven particles. <http://collective-dynamics.eu>
- DiNenno PJ (ed) (2002) *SFPE handbook of fire protection engineering*. National Fire Protection Association
- Knoop VL, Daamen W (eds) (2016) *Traffic and granular flow '15*. Springer, Berlin/Heidelberg (see also previous issues of this conference series)
- Moeslund TB, Hilton A, Krüger V, Sigal L (eds) (2011) *Visual analysis of humans – looking at people*. Springer, London
- Online data archive of experiments studying the dynamics of pedestrians. <http://ped.fz-juelich.de/dataarchive>
- Predtechenskii VM, Milinskii AI (1978) *Planning for foot traffic flow in buildings*, Amerint Publishing, New Delhi
- Schadschneider A, Chowdhury D, Nishinari K (2010) *Stochastic transport in complex systems: from molecules to vehicles*. Elsevier, Amsterdam
- Schreckenberg M, Sharma SD (eds) (2002) *Pedestrian and evacuation dynamics*. Springer, Berlin
- Song W, Ma J, Fu L (eds) *Pedestrian and evacuation dynamics 2016*. Available from <http://collective-dynamics.eu/index.php/cod/article/view/A11> (see also previous issues of this conference series)
- Still K (2013) *Introduction to Crowd Science*. CRC Press, Boca Raton
- Timmermans H (ed) (2009) *Pedestrian behavior – models, data collection and applications*. Emerald, Bingley
- Tubbs JS, Meacham BJ (2007) *Egress design solution – a guide to evacuation and crowd management planning*. Wiley, Hoboken

Natural Disasters and Evacuations as a Communication and Social Phenomenon

Douglas Goudie

School of Earth and Environmental Sciences,
James Cook University, Australian Centre for
Disaster Studies, Townsville, QLD, Australia

Article Outline

Glossary

Definition of the Subject

Introduction

Concepts, Language, and Mathematically

Modeling the Propensity to Evacuate

Mathematical Modeling

Effective Risk Communication

Integrating Theory and Implementation

Institutional Barriers to Greater Community Self-Help

Experiences and Lessons: Some Case Studies

Discussion

Bibliography

Glossary

Community A group of neighbors or people with a commonality of association and generally defined by location, shared experience, or function (Lichtenberg and Maclean 1991).

Community empowerment Internally and externally nurtures a community to accept that residents live in a hazard zone, and they choose to do things as a group to maximize their safety.

Community safety group Existing community groups (such as neighborhood watch) and individuals, working with formal response organizations, form a coherent affiliation in and near a hazard zone, to help maximize safety and care for all community members.

Disaster The interface between an extreme physical event and a vulnerable human population (Skertchly and Skertchly 2000).

Disaster lead time The time taken from first detection of a natural disaster threat to the likely time of impact on humans or human structures.

Disaster threat A natural extreme event which may impact on a community.

Effective risk communication That which motivates people to maximize their own safety through their own and collective actions.

Emergency An actual or imminent event which endangers or threatens to endanger life, property, or the environment and which requires a significant and coordinated response (King et al. 2006).

Evacuation People relocating to safely escape hazardous disaster impacts. To move from a high danger zone to relative safety.

Hazard A source of potential harm or a situation with a potential to cause loss. A situation or condition with potential for loss or harm to the community or environment (King and Goudie 2006). Hazard is synonymous with “source of risk” (EMA 2000).

Hazard zone Defined geographic areas which may be subject to a natural disaster impact of flood, bushfire, storm surge, destructive winds, earthquake, landslide, or damaging hail. Hazard zones include major accident sites, including industrial, transport, or mining precincts, or biological or terrorist threat or impact, or from-source predicted area(s) of pandemic spread.

Mitigation Any efforts taken which may reduce the impact of a threat.

Prevention Measures to eliminate or reduce the incidence or severity of emergencies (King et al. 2006).

Ramp-up preparations The final set of preparations and precautionary evacuations taken ahead of a forecast disaster impact. This includes earlier final actions than precipitated by formal organizations.

Risk treatment options Measures that modify the characteristics of hazards, communities, and environments to reduce risk, e.g., prevention, preparedness, response, and recovery (King et al. 2006).

Vulnerability Comprises “resilience” and “susceptibility.” “Resilience” is related to “existing controls” and the capacity to reduce or sustain harm. “Susceptibility” is related to “exposure” (EMA 2000).

Definition of the Subject

This article intends to show how system and complexity science can contribute to an understanding and improvement of evacuation processes, especially considering the roles of engaged communities at risk, the concepts of community self-help, and clear communication about local threats and remedies. Evacuation ahead of a natural disaster impact is perhaps among the more extreme actions private citizens are encouraged to make by authorities to help them “get safe and stay safe” – the core goal of all disaster risk management and effective risk communication.

This article shows researchers in Complexity and Systems Science (CSS) a social science approach to maximize effective and precautionary evacuation, maximize safety, minimize loss, and speed full recovery. The computational and analytical modeling tools of CSS may be considered to apply to a complex interaction of community awareness and inclination to accept the reality of a natural disaster threat, along with achieving background and final preparations to maximize safety and recovery from a natural disaster impact. This article may stimulate CSS researchers to develop detailed models of the complex systems and complexity of melding information from weather bureaus and disaster managers, via contacts and intervening media to communities at risk, with the shared social goal of maximizing safety. This social science task requires cross-disciplinary approaches of respect and response.

The old disaster management model lacked the predictive and rapid communication systems now available and developing in disaster predictive

models (such as flood maps). An approach to modeling the great complexity of human behavior responding to threat is provided. Such a model must include people’s prior knowledge of a threat type, and consider such fine detail as the overarching language used in a country with threat zones, and the dominant languages of all under threat. It is hoped this article stimulates CSS models to further engage in this social good of helping people get safe and stay safe through natural disasters by providing predictive tool to authorities to better inform and encourage those at risk to action, including the possible need for precautionary evacuations ahead of a predicted impact.

Disaster management in Australia, and increasingly, globally, is focused on mitigation as part of a “threat continuum,” from acceptance that some locations are vulnerable to a hazard impact to recovery (COAG 2004). Emergency warnings and a possible need to evacuate are embedded as “spikes” on that continuum. Thus, this article stresses the importance of developing ways, incentives, to mobilize aware at-risk community members to precautionary self-evacuation. For this to happen, people need to know and internalize the reality that they are in a hazard zone.

Thus, in the cost-effective philosophy of engendering self-help, the process of understanding the complexity of achieving the shared social goal in maximizing safety and minimizing loss is to engender creation of empowered communities with a high motivation for safety-oriented and precautionary action. This is likely to lead to minimized loss and disruption and maximized recovery. This article details many elements of that process and invites detailed development of the sustainability implementation research (SIR) to achieve that goal through CSS.

To model the path to collective safety, the complexity of the dynamics at play need to be clarified: impact preparedness, including possible evacuation, is a communication and social issue.

This article demonstrates that mapping hazard zones (where natural hazards *can* impact) and sharing with those at risk is important for those at risk to internalize the fact they may be at risk as a first psychological step in taking responsibility to do things to minimize that risk. Involving the

community in the acceptance of threat and needed action, developing a safety-oriented social norm feeds back to individuals [SoEaESI] and households.

[SoEaESI] Each Under “Preparations”

Evacuation modeling is needed only for those whose homes may be at real threat of a disaster impact. For those living in a *hazard zone* (e.g., Figs. 2, 7, 11, and 15), a fully informed community, who have internalized the reality of the threat and have worked for maximum background preparation and have mechanisms to receive alerts and warnings of a looming threat, a community predisposed to precautionary evacuations will result. Evacuation modeling is also needed for those who may manage the travel routes or receive evacuees.

Capturing this complexity is the challenge for modelers. Evacuation is about hazard zone residents actively monitoring a looming threat via refined communication channels detailed in this article, within a developed social predisposition to act. Some examples are given. For consideration by scientists and students internationally, this article introduces the Communication Safety Triangle (CST) and the seven steps to community safety on the preparedness continuum, within the new research frame of sustainability implementation research (SIR).

Introduction

The purpose of this article is to share with modelers and complexity and system scientists the social and communication issues of modeling effective safety strategies to a natural disaster threat. It is hoped, through the approaches and processes described in this article, that modelers will more clearly link physical threats with warnings and community engagement.

This article first looks at the definitions and language used in risk communication and effective warnings, leading to informal and formal evacuations, and then considers some Australian policies relating to emergency management. Theories related to risk communication are presented, with examples of evacuation issues provided from

indigenous communities and from non-English-speaking households. The needed conceptual shift to self-help is placed within the larger theoretical frame of paradigms and paradigm shifts.

An example including residents to internalize threats is given, followed by a more general example of transport evacuation.

A discussion of international evacuation issues precedes a broader view of some of the institutional barriers which may restrict the uptake of the seven step approach to an aware, informed community, relying on accurate information and choosing to self-evacuate as a precaution. These issues are discussed, considering effective ways of allowing people to know that they are at risk so they are inclined to evacuate themselves, as a practice. These approaches can be used or tested by other scholars. Recommendations and a summary of the key issues to maximize voluntary and safe evacuations finish this article.

This article was first published by Springer in 2008 and upgraded in 2013. The literature continues to dwell on motivating people to achieve the seven steps to community safety developed in this article (Fig. 8), with the United Nations and others making increasing internet access important both in hazard zone mapping and ways of sharing how to help people get safe and stay safe (i.e., Carlson et al. 2013).

Apart from some text revision, a main change in this upgraded 2013 article is more thought on the integration of theory and practice with decision-makers and knowledge holders in sections “[Integrating Theory and Implementation](#)” and “[Institutional Barriers to Greater Community Self-Help](#)”. There I give greater attention to the relationship of those in disaster management senior professional roles and their inclination or otherwise to use the results of rapidly expanding technological improvements in *two-way* communication in their organizations and interactively with those at risk.

This is important because such use may diminish the power and control of their organization and of their own hard-won status. A similar systems problem, perhaps a fatal flaw, is the possible disinclination of some disaster management organizations to use evidence-based research, which largely says: empower those at risk by providing

them with real-time condition updates at a very detailed and localized level. This does not necessarily sit well with a top-down, command, and control organizational model.

Some Key Evacuation Issues

In an era of increasing social self-help (Boura 1998; King et al. 2006) and community empowerment (COAG 2004; EMA 2000, 2002a), a part of global efforts to embrace ecologically sustainable development (Bruntland 1988; Loudness 1977) is to increasingly see evacuation as a social phenomenon.

This article provides a conceptual frame, the Communication Safety Triangle (CST), which includes responsible media telling people at risk what they need to know and *seven steps to community safety* (7SCS). The 7SCS help guide emergency managers and modelers to treat the possible need for evacuation as a decision-making process where the community should be aware of the potential threats and receive clear, detailed, and reliable information on the possible need to evacuate, so most residents in a high impact zone self-evacuate in a precautionary way, as a practice. Examples provided in this article illustrate this new, *sustainability implementation research* (SIR) approach to disaster management.

Warnings precede a perceived need to evacuate. In the USA, the need for an integrative approach to warnings is identified: "There is a major need for better coordination among the warning providers, more effective delivery mechanisms, better education of those at risk, and new ways for building partnerships among the many public and private groups involved" (Munro 1995).

Sustainability implementation is the way to achieve a sustainable future. Disaster management, effective risk communication, and community self-help provide a stark and comprehensible example of what sustainability implementation means and how it will benefit societies and will help channel us into a safer and more sustainable future. Within this approach, scientists and students become major agents for sustainability implementation, as models can illustrate the needed paths to achievement.

Evacuation Overview

There are three types of disasters which may require evacuation: human-induced and natural disasters with and without lead times. This article is focused on precautionary evacuations ahead of natural disasters with lead times (Table 1). The research and conceptual frames draw on and are applicable to all communities in hazard zones. Defining, modeling, and effectively communicating to at-risk residents and travelers about specific geographic hazards and safety-oriented behaviors are core elements of successful precautionary evacuations.

Within Table 1, evacuation may be a response to an emerging hazard of indicated intensity, direction, and speed. This article advocates development of informed and directed communities which will actively respond to a communicated threat, with the vulnerable moving themselves early to places of safety or being helped by other community members to move.

Concepts, Language, and Mathematically Modeling the Propensity to Evacuate

Precautionary self-evacuation pivots on knowing who is at risk. With the help of researchers and modelers, planning and community involvement can prepare people and their valuables to minimize loss. Very public hazard maps will help people internalize that they are in a hazard zone and what they should do. The concept is not new or alarming: every public building has emergency exit routes marked, with all that implies. Air flight comes with the mandatory emergency preparation presentation. Minute or recurrent risks to where we live or travel to should be no less public, nor more alarming than a fire drill in a public building.

Hazard Types: Little or Considerable Warning Times

Hazards may or may not have lead times (Table 1). Sudden-onset threats include tsunami, earthquake, major eruption, major toxic spill or discharge, mine disaster, and terrorist

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Table 1 Evacuation decision matrix – short to long warning times

Evacuation decision matrix and evacuation around <i>sudden-onset impact</i>			
Hazard	Landslide	Earthquake	Tsunami
Possible safety-oriented response	Stay in strong structure. Move across slope as soon as possible	Get into the open or shelter under strong structure	Immediately flee to higher ground
Precautionary evacuation (PE) decisions with lead time – signaled threats			
Considerations for evacuation decision	Hazard		
	Destructive wind/cyclone	Fire	Flood
1. Vulnerability of present environment	If likely to be in a storm surge, must PE; if shelter weak, must PE	House material, surrounds, water available. If poor, PE	May be inundated = PE; may be cut off, consider PE
2. If 1 OK: vulnerability of individuals (e.g., weak)	PE first	Asthmatics and less able: PE	Judgments of flood height. If in any doubt, PE
3a. Distance to safe shelter	The further, PE earlier	The further, PE earlier	The further, PE earlier
3b. Safety along exit route	Know in preparation	Know in preparation	Know in preparation
3c. Means of travel	Reliability and suitability	Reliability and suitability	Reliability and suitability
4. Community cohesion	Help available	Help available	Help available

threat or attack. Signaled threats include cyclone, flood, fire, and destructive winds. This article focuses on disasters with sufficient warning periods to be able to evacuate the vulnerable away from the predicted worst impact areas. Table 1 that with sudden-onset impacts, sheltering to survive the first few minutes is critical, and then moving to open, stable ground is important to avoid further aftershocks or landslide. Always take direction from local authorities. For natural disasters with lead times, Table 1 shows that if you will clearly be safe where you are, stay put, and prepare as best you can. If, in the worst case, you may not be safe, move early (Table 1). The vulnerability of individuals needs to be considered. In a bushfire, for instance, the general advice is if the property is well prepared (Boura 1998), stay and defend. If, however, you may not be able to cope with the psychological terror of staying through the fire front, fully prepare your property, and then leave early (Table 1). If the exit route may be blocked by flood waters, or by dense smoke or fire, evacuation needs to precede that obstruction. In the “background” phase of community disaster preparation, all such possible obstacles to a clear escape route (including

gridlock congestion) need to be factored in to the timing of precautionary evacuation (Table 1).

Finally, as seen in the Woodgate Beach example of section “[Experiences and Lessons: Some Case Studies](#)” the level of community support in ensuring all residents are aware of and prepared for the “ramp-up” phase of a possible disaster impact, and receive the warning as soon as possible, the able-bodied will help the less able to get out of harm’s way early in the threat period, as a safety-oriented practice, minimizing the demands of the formal response groups as the likelihood of impact increases.

All effective communication involves sending signals and having them received and processed and then incorporated into the receiver’s world view (EMA 2002b; Gurmankin et al. 2004; Rohrmann 2000; Skertchly and Skertchly 2000; Sorenson and Mileti 1991). Effective risk communication (Baram 1991; COAG 2004), section “[Effective Risk Communication](#),” motivates people to act to maximize their own safety. This may not happen if people have not internalized that the threat is real. They are in denial. Alternatively, people may be ignorant of the threat, or why it should be taken seriously. Salter (Rounsefell 1992) categorizes ignorance from pure ignorance to acts of ignoring.

Disaster Definitions

A disaster may be seen as a negative impact of a hazard on a community as a measure of vulnerability. The language of disaster mitigation evolved and is increasingly practiced since the late 1990s (i.e., COAG 2004; Yeo 2002). Risk is seen as a function of probability and consequence, related to exposure and the level of force embedded in the threatening hazard.

Boughton (1992) points out that natural disasters are usually extremely rare for the individuals concerned but they can cause massive impacts. Because Australia is so vast, overall there are reasonably frequent natural disasters. However, in most locations, they are rare indeed. Boughton (1992) argues that a “natural disaster” is a natural event in which the community life is seriously and traumatically disrupted. Embracing the way forward with “structural mitigation,” “. . . a key step in preventing natural disasters is to prevent building damage” (Boughton 1992). Like “disasters,” “community” is difficult to define (Stern and Fineberg 1996). As developed in section “[Integrating Theory and Implementation](#),” “community” is the collection of people in a close geographic area, particularly focused on near neighbors and supportive friends of those in or near a known hazard zone.

What to Communicate

Having the right words or approaches in place *as policy* does not automatically guarantee community safety-oriented responses to disruptive warnings. Since 1989, the approach to cope with disasters has been *prevention, preparedness, response, and recovery* training courses. This helps focus all concerned on the temporal sequence. The language could perhaps be refined to talk about acceptance that a threat exists, background, then “ramp-up” (final) preparations to safe shelter, impact, and then orderly return and recovery. Classically, “response” was seen as the final, near impact flurry of “lockdown” activity (Leibovitz 2003).

Deeper than language is our conceptual frame (section “[Integrating Theory and Implementation](#)”). If researchers encouraged disaster managers to move “response” behavior back a day, a

few hours to “precautionary, early response,” many of Lewis’s (Leibovitz 2003) legitimate concerns would be addressed. Some disaster managers may see themselves as dramatic figures in the early impact phase of a disaster, rather than calm, precautionary minimizers of risk. This culture has changed greatly over the last decade (section “[Integrating Theory and Implementation](#)”). For many reasons, there is a clear divide between the US federal response to Cyclone Katrina (2005) and to the California fires (October 2007). This shift from passive to active can be encouraged by modelers. As part of effective risk communication to encourage precautionary impact preparedness, research with remote indigenous communities and recently arrived, non-English speaking refugees (section “[Integrating Theory and Implementation](#)”) showed the importance of accurate, plain English and the use of images.

Yates (Woods and Gabriel 2005) argues that mitigation efforts need to be refined to make sure they are focused on issues from the relevant local communities. Much of the problem of nonresponse seems that “the message” to take care does not effectively get through to the target (Paton 2003). It has to do with communication, signals sent, signals received, and their interpretation.

The types of received and interpreted messages are explored specifically in section “[Integrating Theory and Implementation](#),” which develops the intellectual foundation to more fully understand the semiotics of risk communication. The following outlines the concepts and relevance of semiotics.

Semiotics is *the study of signs* including words, sounds, and such things as “body language.” By the “message conveying” principles of semiotics, we can understand how various authors on disasters and evacuations approach the topic and which cultural signs and symbols they manipulate. For example, a sign showing a person running upslope ahead of an exaggerated tsunami wave contains all the preferred imagery to tell people what to do in a tsunami warning. Images, as seen for locating oneself in airports or finding evacuation stairs, do not use words.

The next section lays the foundations for a mathematical modeling of propensity for householder self-evacuation, centered on knowledge and refined communications acting on residents in hazard zones who are predisposed to act in their own safety.

Mathematical Modeling

Foundations of the Household Safety Preparedness and Action Index: The HSPA Index

This section does not consider refined algorithms to define how successfully people may respond to *Authority's Perceived Need to Evacuate (APNE)*. This section considers the form and ingredients of how this could be done, the elements and their possible interactions to predict the success of APNE. Unfortunately, in the previously cited Canberra 2003 bushfires, the APNE did not appear universally, nor was it properly conveyed to the residents in dire risk.

Rohrmann (Renn and Rohrmann 2000) has clearly mapped a process from the warning signal to the hoped-for response. The information and processing flow to internalized inferences is called intuitive heuristics (Renn and Levine 1991). Renn and Rohrmann (Renn and Levine 1991) basically argue that people process probabilistic information on the likelihood of them needing to move, aligned and convergent with researchers like Thompson (2002), hence the need for the seven-step process (Fig. 8) to trigger a paradigm shift for people in danger zones.

As seen in formula 1, there are finite evacuation triggers (ET), ranging from no lead-time LT_0 , such as an earthquake, to an LT of days (LT_{xd}), such as a cyclone. So, a key component of a successful response to an APNE is the time in which to disseminate the warning (LT_{0-xd} ; Table 1). Also, impact severity (IS) is central to safety and loss.

As detailed in section “Integrating Theory and Implementation” and Fig. 8, *the seven steps to community safety*, there is a convincing body of research which indicated that people’s propensity to respond (PPR) to an emergency warning (whether that response be to finalize preparation

to stay and defend the property or evacuate as a precaution) is strongly linked to their Acceptance they are In a Hazard Zone (AIHZ).

Community resilience (linkage and inter-support and communication – CR) is also important, along with a potential evacuee’s knowledge base of the hazard and safety-maximizing behavior (KB). Thinking through to recovery (TR) is also important. As per points 6 and 7 in the 7SCS (Fig. 8), the medium of warning delivery and the quality of the warning information (Medium and Message, MM) are factors in the response of those under threat. Overall, Sorenson and Mileti (Skertchly and Skertchly 2000) believe that increased credibility of the warning source means the specific warning will be more effective. They also provide evidence that the electronic mass media produced the most believable public warnings. This underlines why development of formal links between all types and levels of media information about hazards and preparation from the weather bureau and emergency managers is so important. Having formal links with the media, coupled with web-delivered “real-terrain” simulations of the hazard, will help produce a powerful and effective “active warning” regime.

Issues of exit route(s) (ER), possible travel mode (PTM), may be crucial. A normally free-flowing exit route may be blocked by the disaster itself or others trying to flee. This may involve accidents. People’s psychological and physical states of well-being (WB) will also affect potential effective evacuation decisions and accomplishment. Authors like Rohrmann (Renn and Rohrmann 2000) have attempted to stylistically model some of this complexity.

Many of the earlier recognized impediments to full preparedness, like unrealistic optimism (Handmer 2001), are incorporated into one overarching factor of formula 1: people’s propensity to respond (PPR), including the centrally important feature of individual households fully accepting they are in a hazard zone and that they need to engage in background preparation, “listen up” around a hazard threat, and undertake their own, precautionary final preparations.

Cutter et al. (2003) developed 11 factors indicating vulnerability caused by a major disaster, using

principal component analysis within factor analysis. The resultant Social Vulnerability Index (SoVI) necessarily weights the 11 components according to the percentage of variation of analysis of counties on the USA as to resident's vulnerability. Personal wealth, age, and housing density head up the contributions of the 11 - variables used to determine the SoVI of US counties. Cutter's SoVI contributes to the likelihood that proper preparations, including for a timely evacuation, would occur, either at a broad regional or more individual level. This SoIV is included in the HPSAI below. Its weighting remains to be tested. A final issue overarching many others is institutional barriers to change (IB), detailed in section "[Institutional Barriers to Greater Community Self-Help](#)." The support or otherwise of the media can play a critical role in effective crisis communication – most messages are likely to be through the media, so the media is part of the systems complexity blend: media support (MS).

From the weight of SIR and the conceptual framework described in this article, the following generic formula expressing the above factors likely to impact on householder propensity for precautionary evacuations may entice other, more theoretical researchers, to develop the needed algorithms. What follows is purely a design base for others to develop predictive modeling of *People's Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay)* PPCSE is better named the *Household Preparedness and Safety Actions Index: the HPSA Index*. The following is a synthesis of Bayesian logic (Hoeting et al. 1999) and eigenvalues.

People's Propensity and Capacity to Successfully Evacuate (PPCSE) or Household Preparedness and Safety Actions Index (HPSAI)

$$\begin{aligned} \text{HPSAI} \propto & f(\text{PPR})(\text{AIHZ})(\text{LT}_{0-xd}) \\ & (\text{APNE})(\text{ET})(\text{IS})(\text{AIHZ})(\text{CR})(\text{TR})(\text{MM})(\text{ER}) \\ & (\text{PTM})(\text{WB})(\text{SoIV})(\text{IB})(\text{MS}) \end{aligned} \quad (1)$$

Formula 1 anticipates constants to "weigh" each factor according to its contributing importance on

the resultant HPSAI (see Heller et al. 2005, p. 384 for detail of the Bayesian modeling approach).

Resultant safety (*safety* being a combination of successfully evacuating to a safe place or sheltering within the impact zone in a safe place) of those under threat will be greatest (Fig. 1) where the factors of formula 1 tend to intersect in the positive quadrant of an eigen plane (Fig. 1) (Woods and Gabriel 2005). Equally, outcomes of loss will occur with, for example, short lead times, insufficient warnings of evacuation triggers, or lack of belief that residents or travelers are in a hazard zone. A journal search showed few links between disasters and eigenvalues. An exception is Fowler et al. (2007). With global warming and increased populations increasingly encroaching into obvious hazard zones, modeling to influence planning and to maximize HPSAI must be a major growth industry.

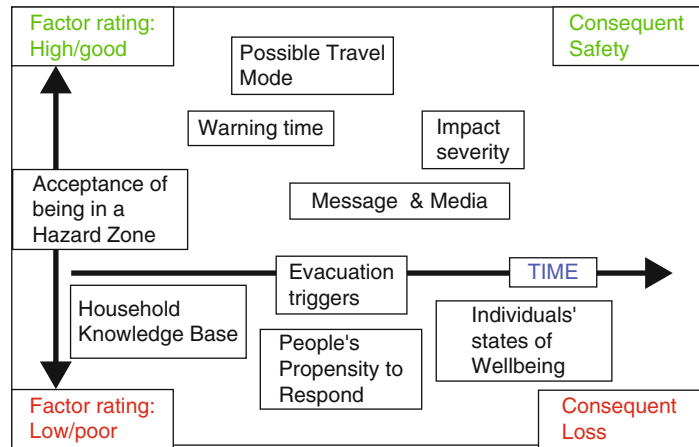
Calibrating the Model: The Greek Fires, August '07

Kikuchi and Nakamori (2007) or California in late October '07 (BBC 2005) shows that if various of the above variables or likely variable clusters or eigenvalues from factor analysis have low values, people will not be placed in a strong, precautionary position of safety. This was true of fire-threatened residents in Canberra in 2003 (<http://www.abc.net.au/canberra/bushfires/>), all ultimately leading back to the mathematical weight which needs to be given to each point on in the 7SCS: especially accepting that the threat is real (and possibly imminent).

Once the above-described preparedness paradigm is internalized, it is difficult to experience any disasters news coverage without feeling some despair: cars submerged or floating in flood waters, people fleeing for their lives ahead of imminent immolation, and any reports of lost lives through natural disasters with lead times of hours or longer. There are few excuses for this to remain so. In most parts of the developed world, there are prevailing technologies to detect and rapidly convey threat. Researchers are well placed to conceive, trial, and link the threat information directly to prepared populations, to close the gap between knowing and acting to maximize safety,

Natural Disasters and Evacuations as a Communication and Social Phenomenon,

Fig. 1 Factors in the eigen plane determining *Household Preparedness and Safety Action Index*



as a social exercise of community engagement and communication refinement.

Cyclone Larry (King and Goudie 2006) showed that, with general community acceptance of the reality of the threat, good warnings and a lead time of about 20 h, combined with well-developed disaster management and media cooperation in getting the needed messages to people well primed to the 7SCS, there was no loss of life, although about 5000 people were directly subjected to destructive winds of a category 4/5 cyclone. Residents of the impact zone generally followed the 7SCS, supporting the general approach of formula 1: the community was prepared and acted in a precautionary way to maximize their own safety.

The above model design offers elements/factors to maximize *People's Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay)*, the *Household Safety Preparedness, and Action Index (HSPAI)*, so strengthening a paradigm among agencies and hazard-zone residents of “self-help” and “community engagement” approaches, less passive than earlier approaches used in considering people’s vulnerability. This generic approach is supported by model testing by Schadschneider et al. (Evacuation Dynamics: Empirical Results, Modeling and Applications).

The Australian examples detailed in this article, including the national shift to mitigation (COAG 2004), best illustrated in the Woodgate Beach example (section “[Experiences and Lessons: Some Case Studies](#)”), show that the conceptual frame and thus variables chosen to mathematically model evacuation propensity and

capacity are more important than the results of any model using “passive” variables or variable clusters. The advantages and means of achieving this shift from community passivity to partnerships, place- and people-based community empowerment, and self-help form the basis of this article. The above mathematical approach may help validate and guide this necessary conceptual shift.

The next section considers policy, laws, concepts, and possible safety-oriented actions that words, images, and stories convey to vulnerable residents in hazard zones. In Australia, policy and legislative requirements are indicated, which aim to responsibly maximize community safety (COAG 2004; ECAP 1997; EMA 2000, 2002a, b, c; Handmer 2001; Phillips 1994). Some new concepts, epitomized by new language phrases (with their embedded meanings), are discussed in section “[Integrating Theory and Implementation](#),” especially *community safety groups*. New phrases include “social burn-offs” and “practice evacuations”; like “the Communication Safety Triangle” and “seven steps to community safety,” it is hoped these expressions enter local, state and national lexicons and modeling in disaster management.

Effective Risk Communication

Policy and Laws on Hazard Preparedness and Evacuation

All risk communication operates within government policies and legislation. Forced evacuation

may be normal in some jurisdictions for all threats and outlawed in others. Policy may encourage passivity, or be energetically proactive in achieving the 7SCS, both in urban planning and building materials and community engagement. In Australia, for instance, there are many federal and regional guides to urban growth, (e.g., AMCORD 1995). A new urban paradigm is, perhaps, best expressed as: “Think globally, act locally, respond personally” (AMCORD 1995 pnp; Australian Bushfire CRC 2005, p. 1). This paradigm shift can easily embrace and be informed by computer modeling of individual’s behavior, demonstrated in recent Springer’s Journal of Systems Science and Complexity articles, such as Kikuchi and Nakamori (Zamecka and Buchanan 2000). Agent model analysis explores effects of interaction and environment on individual performance (Kikuchi and Nakamori 2007).

From all decisions of urban development being in the hands of “experts,” government policy increasingly requires a dialog with local residents, through public participation. The first barrier to ecologically sustainable urban development is “. . . belief systems – doubting a problem exists, or supporting the status quo. [Solutions include] . . . consciousness raising campaigns, public participation in decision-making, demonstration projects and incentives and disincentives” (Rohrmann 2000, p. 46). This is the KB of formula 1.

This article stresses that response models, authorities, and all residents are constrained (and empowered) by national and regional laws and policies. Research can only work within a prevailing frame and will be welcomed by or influence prevailing policy. It is thus useful to have clear goals to achieve social or environmental “good” (COAG 2004; Heller et al. 2005; Phillips 1994). This requires greater community involvement in mitigation and responsible media helping to mitigate disaster impacts (MS of formula 1). In Australia, the national radio broadcaster has a binding agreement with the weather bureau and disaster managers to read issued warnings verbatim and report and engage community members in safety oriented behavior, by providing relevant

information. Local community media are often willing to develop the same responsible functions.

Risk Communication’s Core Goals

Risk communicators advise “individuals and communities to respond appropriately to a threat in order to reduce the risk of death, injury, property loss and damage” (Handmer 2000). Risk communicators and disaster managers need to formally work closely with the media to maximize social benefit.

Although it may be intended that information flows and is received accurately, that desired behavior will result, and that only communication techniques are important (Kasperson and Stallen 1991), humans tend to be irrational and optimistic and only hear what they want to hear. It is not what our message *is*, but what, if anything, the listener *does* with our message. To have any chance of “success,” information needs to have meaning which is shared between those who construct and send the warning and those for whom the warning is meant to inform and motivate to action.

“Action statements” (what the at-risk person, family, or community should actually do to minimize damage) are seen as central to the whole purpose of risk communication (Salter 1992). Kasperson and Stallen (1991), along with Salter (Rounsefell 1992; Salter 1992) and others, detail risk communication messages in terms of content, clarity, understandability, consistency, relevance, accuracy, certainty, frequency, channel, credibility, public participation, ethnicity, age, gender, roles, responsibility, elements, sequencing, synopsis, prognosis, location, action, warning timing, and action statements. It is not a case of saying: “a category three cyclone will pass over Smithfield.” It is more a case of making sure the members of Smithfield hear that message and feel moved to and competent to take well-understood personal risk-minimizing actions.

Knowledge, Self, World Views, and Messages

For precautionary evacuations to be successful, the seven steps to safety (section “[Integrating Theory and Implementation](#)”) have occurred, or people have been coerced by authorities, or they have seen others depart and felt insecure, so they

leave as a follower of the “innovation” of the “norm,” precautionary evacuation. This occurred in the desktop exercise in an isolated settlement in March 2007, detailed in section “[Experiences and Lessons: Some Case Studies](#).”

It is now difficult to understand why people needed prompting to take evasive action against the forecast Brisbane floods in 1974, but many ignored the prompts. Authorities believed houses and roads would be flooded, so people should finalize travel or evacuation early (Chamberlain et al. 1981a). For those in the flood zone, 88% of a later survey sample reported evacuating their home. Some took this step very early, but 67% of respondents only made preparations immediately before leaving home. Twelve percent only left on foot or in boats *after waters entered the main living areas of their homes*. Almost 22% of respondents said they made no preparations, mainly because the threat was not recognized in time (Chamberlain et al. 1981a). This well-documented “poor” crisis communication can help calibrate formula 1.

A key safety message surrounding floods is not to enter flood waters. Figure 2 is one example of using images to help convey the reality of a threat and the pitfalls of belated action.

Effectively Conveying Meaning

Because language is pivotal to acceptance of risk and conveyed warnings of need for evacuations, it is important that all participants reasonably agree on word meanings. This section starts in the comfort zone of simple definitions, then considers

semiotics and core issues of world views (paradigms), and finishes with the uneasy realities of our knowledge base, our epistemological orientation. Section “[Integrating Theory and Implementation](#)” shows that some cultures do not carry the background cultural experience to absorb the meaning of cyclone or bushfire – there is no shared experience or “stories” of the power of these extreme forces of nature. Some of the underlying knowledge foundations of context, intent, and behavioral motivation need to be considered – how humans construct, transfer, acquire, and use knowledge.

Imparting meteorological knowledge, then warnings, to target audiences to engender safety-oriented responses is a complex exercise in social empowerment, explored in section “[Experiences and Lessons: Some Case Studies](#).” As information promulgators, information and warning sources should understand a little of *perception* (the raw data from the outside world entering an organism via any of the five senses), *cognition* (internal processing, analyzing, information storage, and processing), *attitudes* (how we think and feel about particular issues, implying a predisposition to specific action), *language use*, and links to *behavior*.

An intellectual framework to risk communication is provided by Rohrmann (Renn and Rohrmann 2000). It considers “the message features”: the recipient features, social influences, and context which influence individual risk assessment and management, including preventative action. Rohrmann and Handmers’ publications (Handmer 1992, 2000, 2001; Renn and Rohrmann 2000) inform



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 2 What flood-threatened residents need to consider (Photos courtesy Townsville City Council)

the Communication Safety Triangle and *the seven steps to community safety* provided in this article.

Philosophy for Policy Review: Crying Wolf or Worse, Applying the Precautionary Principle

From the 1990s, a strong issue of debate in risk communication has been “the right to know” (Baram 1991). Some disaster managers wish to avoid false alarms, which may cause “concern fatigue.” This can be seen as an institutional barrier to change (section “[Institutional Barriers to Greater Community Self-Help](#)”). “Avoiding undue alarm” is in conflict with the right to know and the precautionary principle of ESD.

The “precautionary” approach is supported by the Economic Commission for Asia and the Pacific (ECAP), the World Meteorological Organization, and the Red Cross Societies. The alerting of the community and its responsible authorities must begin, at least provisionally, as soon as the existence of a tropical cyclone over the seas bordering the country is known (Dunlap and Van Liere 1978, p. 16). According to ECAP et al. (Dunlap and Van Liere 1978), the warning challenge is less clear for predicted localized downpours and flash flooding how much effort should be taken to warn – what is the message, how do you keep it to the affected area, and what do you want people to do? These questions resonated in Australia after the billion dollar hail damage in Sydney in April 1999 or damaging flash floods in Melbourne in December 2003.

Precautionary Evacuations

Handmer (2000) reports an evacuation of 250,000 Dutch ahead of a flood threat in 1995. Eighty-eight percent of people surveyed in broad post-emergency surveys “believe that evacuation was appropriate” (Handmer 2000, p. 24). In part this may be because of floods experienced 2 years prior. Good skills in dealing with the mass media appear to have helped in the effective precautionary evacuation. The Dutch experience showed a willingness to evacuate again in the future, even though the threatening flood waters did not inundate to the level feared. This compares favorably with the 2005 boat owners’ responses to Cyclone Ingrid in Port Douglas, North Queensland (section

“[Institutional Barriers to Greater Community Self-Help](#)”). Both Handmer and Goudie’s research (Goldammer 2005; Goudie 2004, 2007a; Handmer 1992, 2000; Kitchin 1996; King and Goudie 1998, 2006) show clearly that people would rather practice (make a precautionary evacuation) than incur loss. It was treated as a learning experience. The “boy who cried wolf” argument is not acceptable. This is an important message researchers can test and convey to partnering disaster managers.

Have Clear, Consistent Messages

Part of the *seven steps to community safety* is clear, reliable, explicit languages and images. Salter et al. (Salter 1992) point out that the use of meteorological category systems such as “minor,” “moderate,” or “major” carries unambiguous information about the level of disruption likely from a particular flood. Language used should not be for the convenience of the warning agencies. Its function is to convey clear unambiguous messages to the threatened public.

Use Past Events to “Make the Threat Real”

The purpose of risk communication is to make people perceive the threat as real and to successfully motivate safety-oriented action. Boughton (1992, p. 6) argues that awareness of hazards and disasters can be fostered by “drawing attention to media coverage of hazards in other places.” Images of large-scale floods and evacuation in Holland in 1995 (Handmer 2000) may help prompt a future flood evacuation.

The Changing Politics of Risk Communication

Risk communication is often laden with values and political implications (section “[Institutional Barriers to Greater Community Self-Help](#)”). For instance, in the 1990s in Cairns, North Queensland, it was argued that the reason for not having detailed local cyclone surge inundation maps made available at the corner store level was that such information may have a negative impact on local land prices. This appeared more a political decision than an attempt at effective risk communication. Now such maps are freely available on the local council web site (Cairns City Council 2007).

Sirens or Not

No or short lead-time disasters are a powerful argument for warning sirens to alert people, perhaps at 2 am, to “listen to local media now.” Lives will be saved in future tsunamis and bushfires with the reintroduction of sirens, as insistent triggers to “find out more now.” However, “large numbers of sirens are needed to cover populated areas and to be loud enough to be heard indoors by most people. Sirens are expensive to install and maintain and can only provide limited information” (Munro 1995) p. 33. Fixed public address systems or those on vehicles may be used. People who hear such sirens will be encouraged to tune to local media, and to phone others in the threat zone, to warn them of the alert. Sorenson and Mileti (Sheppard 1986) believe sirens are most effective if used on populations without other ways of receiving the warning. Wider use would appear prudent, especially with short lead-time threats.

Media Roles

As community media moves from “sensationalism” (COAG 2004; QG and QES 2003), the 2007 research by Goudie confirms that community media organizations now want to become responsible conveyors of safety-oriented information to people at risk. After the Canberra fires of 2003, where 300 homes and 4 lives were lost in the nation’s “bush capital,” all local media signed binding agreements with emergency authorities to faithfully convey provided safety information. The Communication Safety Triangle envisages local media and householders drawing detailed local threat information from the Internet, with media conveying that directly to readers, watchers, and listeners. This forms the basis of future “world best practice” risk communication. In the preparation for threat impact, the reliable information will help people make informed decisions, rather than remaining trapped and inactive in uncertainty.

Risk communication is complex, involving many values, predispositions, and distorting lenses. Rohrmann’s (Renn and Rohrmann 2000) fine risk communication explanations show that we may tell the people at risk, but they may not

interpret as intended. Clear, consistent warnings in plain English, with clear images of the threat, showing safety-oriented behavior are needed from reliable sources. Warnings, seen on a continuum of risk and preparation actions, should be able to be discussed and reinforced with information from other sources. This is most likely to produce safety-oriented behavior, with the constrained and clear assistance of the media, as responsible agents for community safety.

Given the previously fraught nature of risk communication, section “[Institutional Barriers to Greater Community Self-Help](#)” provides a conceptual framework, with Australian examples of current community safety theory and implementation, preceding some detailed examples.

Integrating Theory and Implementation

This section introduces a triangular model to best ensure community safety. With disaster managers central, the three elements are the community, local media, and the Internet. Also, there is a continuous spectrum of seven steps, from accepting there is a risk, through early preparation, final (ramp up) preparation, which may include evacuation, surviving the hazard impact and achieving smooth recovery. An exploration of motivation leads to a discussion of “world view,” including some insights into remote indigenous communities (Goudie 2004) and recently arrived non-English speaking immigrants. The conceptual frame and steps are intended to help guide risk communicators and potential modelers through the issues of acceptance, preparations, evacuation, and functional recovery. Within CSS, all these factors have some bearing on consequent behavior.

Changing Values and Roles

In an attempt to understand why we support or ignore certain messages relevant to our safety, and thus facilitate modeling, this section considers the dominant social paradigm and the new environmental paradigm. A paradigm is a coherent world view, “a mental image of social reality that guides expectations in a society” (Dunlap and Van Liere

1978, p. 10; EMAI 1998). Shared paradigms change. The underlying philosophy shift toward disaster mitigation generates a distinct policy move toward safe, self-helping communities. Thus, the goal of disaster managers, effective risk communicators, and computer modelers becomes one of championing or demonstrating the efficacy of self-help techniques and information to and within communities.

World Views Which Accept Responsibilities of Living in a Hazard Zone

“Why do we do what we do?” has long been a central exploration of psychology. Precautionary safety behavior, including self-evacuations, may depend on a person’s world view. Stern et al. (Stern 1992) developed a simple and elegant model to help explain human behavior, starting with a person’s position in a social structure, with constraints and incentives which generate values. Values determine general beliefs, leading to a consistent world view, specific beliefs, and attitudes, which predisposes intent and helps explain behavior (Stern 1992 Fig. 3). Loosely translated, behavior results from a linked sequence starting with the where, when, and to whom of an individuals’ birth, the surrounding circumstances, and “cultural sheath” of their early childhood experiences, leading to acceptance or rebellion against prevailing social norms, determining an evolving world view. Once we have our coherent world view, beliefs and values

give us our “predisposition to act,” our “intent” which precedes action/behavior.

Many things now mold and modify world views, including the media, and normative world views are mutable and changeable. Aligned starkly with our shared and collective biological survival urge, and much concerted efforts by scientists and conservationists for decades, publicizing by people like Al Gore, and the authority of England’s uptake of the 2006 Stern Report, there is now a global paradigm shift to the urgent need for behavioral and technological change to minimize the gross impacts of global warming, pertinent to increased disasters and needed evacuation readiness. Modeling specific hazard zones’ strengths and weaknesses and broadcasting prediction simulations to the Internet and TV will help mobilize those at risk.

Why We Do What We Do

A behavioral model of causality (Wall 2006 Fig. 4) shows relationships between reported attitudes and actual behavior. However, a complex model proposed by Kitchin (Kim and Ball-Rokeach 2006), with the strength of including social and environmental interactions, shows why we do what we do: Kitchin’s model includes a person’s “working and long-term” memory. Internal information is processed within “real-world” context, such as cost (Sorenson and Mileti 1991), which may be processed within the “it can’t happen to me” cognitive frame (Loudness

Natural Disasters and Evacuations as a Communication and Social Phenomenon,
Fig. 3 Stern’s 1995 behavioral explanation model

Behavior is explained by:

1. a person's *position in a social structure*,
2. with *constraints and incentives* as generators of *values*, which lead to
3. general *beliefs*,
4. *world view*,
5. *specific beliefs and attitudes*, generating
6. *intent*, which helps explain
7. *Behavior*.

1977) of subjective reality. These complex but quantifiable attributes will contribute to modeling maximum likelihood to accept, prepare, and act to maximize safety.

Learning from 18 Remote Indigenous Communities

Much of the rest of this section underlines why modelers will need to deal in great detail with “cultural” aspects of massed responses to safety threats.

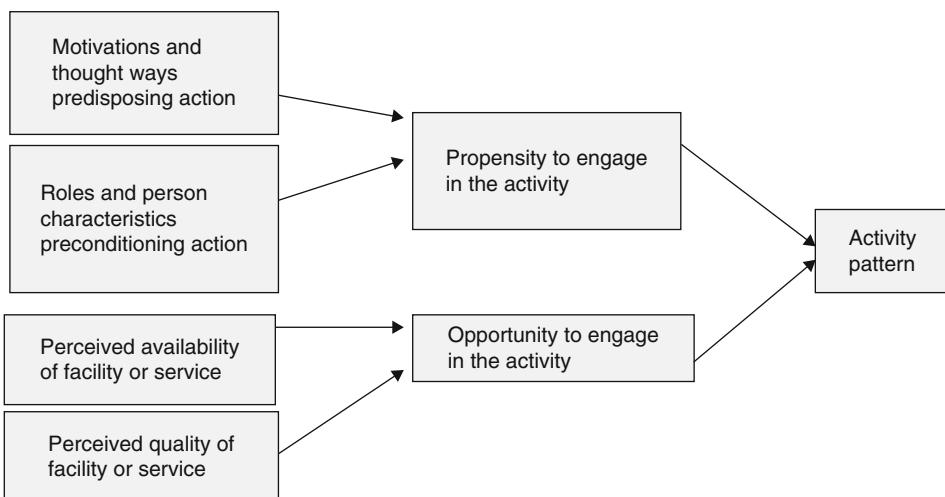
The Stern model (Fig. 3) helps explain why there is such a strong sense of self-help in remote indigenous communities (EMA 2002b; EMAI 1998; Sheppard 1986; Yates 1992). Elders decide responses to threats; there are historic and immediate constraints, generating a value system where community members need to look out for each other.

In many traditional Australian aboriginal stories, most people drown in flood disasters, often as punishment when people do not take care of each other (Goudie 2004, 2007b; NDSI Working Group 2000; Thompson 2002; Utemorrah et al. 2000). If general beliefs embrace self-help, including the felt need to ensure safety, and a resultant intent to achieve community safety, this should lead to safety-centered behavior.

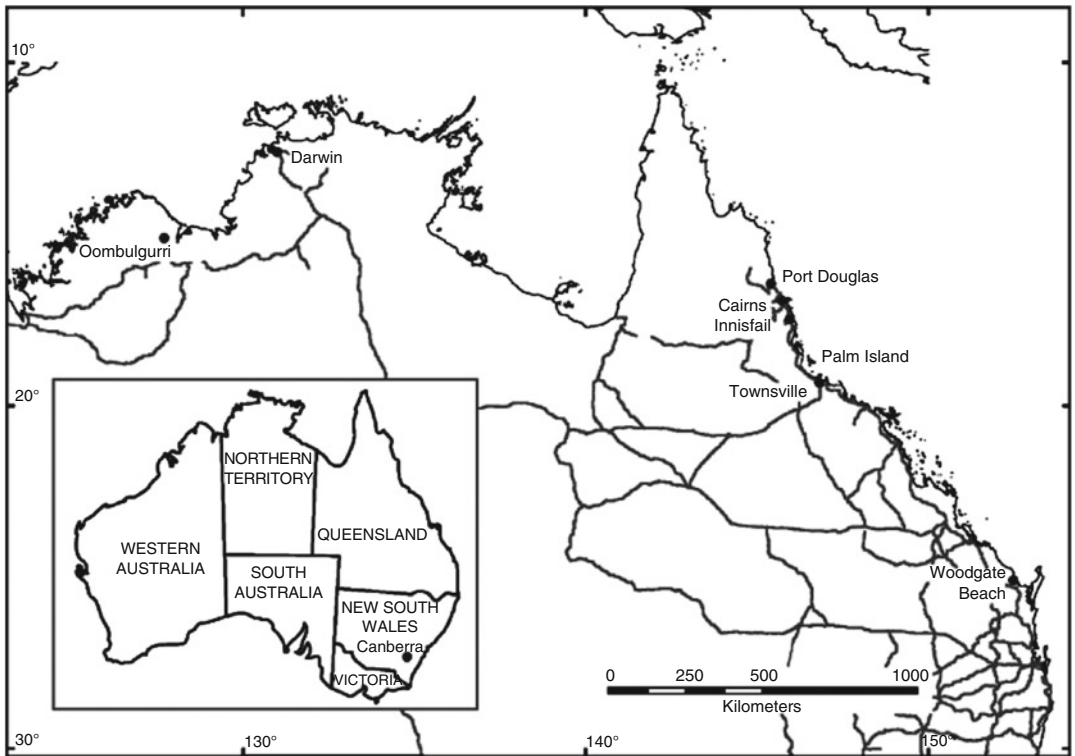
Each community (Goudie 2004 Fig. 5) was not greatly concerned by weather extremes (values), but each relied on and respected their traditional reading of threats and information from the Bureau of Meteorology (the Bureau). The world view is that flooding or worse may happen and that the Bureau will provide adequate warning. Responses may be inhibited by the Social Vulnerability Index (SoVI) of formula 1.

Community Links, Phone Trees, and the Web in Risk Communication for Non-English Speaking Households (NESH)

Many multicultural organizations and focus group meetings in 2005 helped develop the warnings phone tree model for NESH in Fig. 6. NESH rarely listen to the English language media. With about 30,000 NES people arriving in Australia each year (DIMIA 2004), there is federal government recognition of special emergency management needs (EMA 2002c). This way of getting evacuation warnings through was developed once interviews revealed that practically all NESH have a mobile phone and are closely connected with their nearest government-funded multicultural organization and their “community leader,” hence the phone tree. Modeling is of and for the real world, so research like this reported NESH

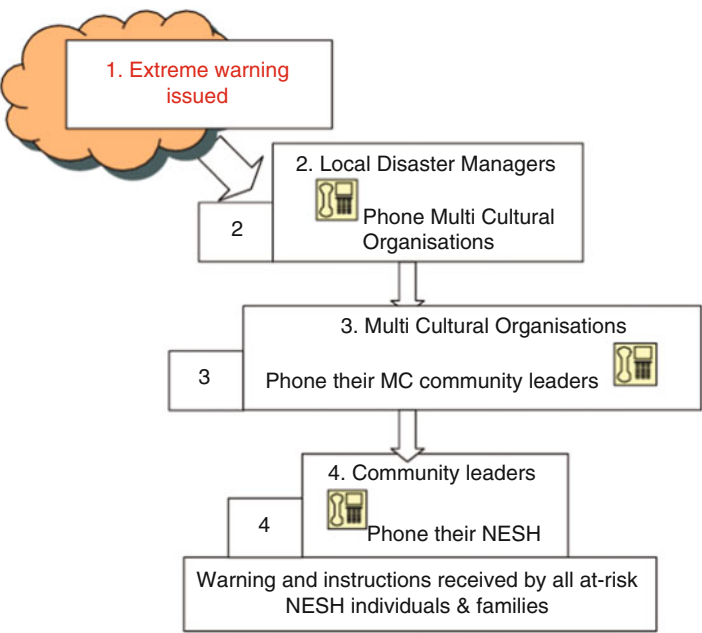


Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 4 Possible determinants of activity patterns (From Wakefield and Elliot 2003)



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 5 Australian study sites

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 6 Multicultural phone tree disaster warnings model



study is needed to see what “complex systems” may be possible. Further recommendations to develop a multilingual warning web site with a simple guide to the seven steps, in relevant languages, to be accessed and used by MCO and NESH training sessions are being considered by authorities. This provides another example of the Internet’s latent role in effective risk communication.

The Australian Northern Territory has cyclone preparedness kits with action guides in eight languages; the NSW Rural Fire Service has information in 27 languages. This shows the importance of modeling all possible communication avenues, encapsulated as Medium and Message: MM in the general formula for Household Preparedness and Safety Actions Index (HPSAI).

A Holistic Approach to Community Risk Management

The seven steps for effective warnings involve community networking, “responsible” media, and increased web use. A “continuum” approach to hazards which may lead to a need to evacuate is important to modelers because it is important to all people in hazard zones: they will be able to access more accurate, detailed, and timely information of any looming threat, assisting emergency managers because a self-help public will decrease demands on formal evacuation, rescue, and property protection (Woods and Gabriel 2005). Recovery will be less arduous (O’Neill 2004) if hazard impact is minimized. Reducing disaster impacts will reduce costs to national benefit. This gives strength to supporting the very demanding goal of developing the model of formula 1.

Risk Managers

Researchers can work with risk managers to achieve “. . . better, timely warnings and advice on safe action during fire events” (Australian Bushfire CRC 2005). With risk managers central, Fig. 7 suggests that enhancing community links (Boura 1998), web information for residents and media outlets, and cooperation of community media with fire managers (Utemorrhah et al. 2000) will more empower householders to embrace self-help

in fire safety. Figure 7, the community safety triangle, with the “seven steps to community safety” (Fig. 8) provides the conceptual frame to enable maximum safety modeling.

Core of Risk Communication for Community Safety Through Natural Disasters

See Figs. 7 and 8.

Building Community Links and Refining Media Delivery Will Change the Household Preparedness and Safety Actions Index

The disaster safety benefits of enhanced community links fit positively with broader social policy (ISR 2007). Stronger community links will help ensure that threat information is easily accessed and shared, and the need to actively self-protect is internalized at the neighborhood level. Residents will benefit by enhanced community links (with such innovations as social burn-offs and Community Safety Groups) in improving their general quality of local social interaction. Changing community resilience (CR of the formula) will then change overall preparedness, so pilot tests of change to CR will show quantifiable changes to the Household Preparedness and Safety Actions Index. Disaster web site managers will be providing a product which is rationalized to reduce national duplication of core information. This will positively change Medium and Message, MM, of the formula. Residents with few English language skills will benefit by having disaster information delivered, via the web, in language and concepts that can be internalized and acted upon.

Handmer (2000) suggests that a flood, for instance, is actually “owned” by the communities at risk. Individuals and organizations within these communities actively seek out information and mobilize their personal networks for action. In this way of looking at the warning process, the warning specialists act as mediators between the threat and the threatened. Local knowledge is used and the whole response process remains focused on safety and loss minimization. For this plausible vision to become fact, local capacity building will need to proceed apace with these

Natural Disasters and Evacuations as a Communication and Social Phenomenon,

Fig. 7 The communication safety triangle



Goal: Maximise safety and recovery, and minimise loss in hazard zones.

1. Encourage those in hazard zones to accept that the risks are real.
2. Help create an aware, informed community, predisposed to safety-oriented action, as a precaution; as a practice.
3. Encourage information-sharing and support among friends, neighbours, family.
4. Provide 'what to do' (action) information, via reliable sources, including web and community media delivered for background and preparation.
5. Encourage people to think right through to impact and recovery.
6. When a threat is closing in, warning messages will clearly convey: *this is real, this is coming at me. I need to make safe where I am, or move early to somewhere much safer. I will not travel during the impact period.*
7. Provide timely, effective threat warnings and fine location and forecast weather detail, and recommended local responses.

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 8

Seven steps to community safety

more formal efforts to inform and maximize community safety.

The importance of “informal” information sources and community links is shown in Table 2, from about 200 people interviewed immediately after Cyclone Larry in 2006 (King and Goudie 1998), showing that about 90% are deeply dependent on social and family support, if only for reassurance.

Finally, for community enhancement within the CST model, there is a national disaster mitigation policy (COAG 2004) <http://www.em.gov.au/Documents/Natural%20Disasters%20in%20Australia%20-%20Review.pdf>: in the statement of the paradigm shift (COAG 2004), p. 13, “Principle 7 – Reform Commitment 7: develop jointly improved national practices in community awareness, education, and warnings which can be tailored to suit State, Territory and local circumstances.”

Community and Social Media

“The Media: Media organizations, particularly public and private radio and television organizations,

have responsibilities in ensuring that timely and appropriate warnings and advice on disasters is broadcast to communities at the request of relevant authorities. They also have a role to play in educating the community about natural disaster issues” (COAG 2004, p. 18). Further, social media is becoming critical to shifting social norms of action under threat. There is a “look over your shoulder” response of evacuation if you know most of your neighbors already have left (Carlson et al. 2013).

The Web

Rationalized web information delivery: One national generic provider and upkeeper of core information, known to all, is recommended, to which state and local government will add their unique detail. A rationalized disaster information web delivery can properly incorporate broad contents and fine (and zoom-in findable) and multi-lingual information.

Simulations of predicted near-term fire or other threat movement based on the above, akin to microscale modeling of cyclone impact forecasts, will help revolutionize the way people react to fire and flood threat. With most households having the Internet, prior to likely power loss residents can draw on enhanced fine detail of the current fire situation, to assist in their actual decision-making. Before Cyclone Larry struck, residents with web

access to the Bureau copied cyclone forecast maps and distributed them to neighbors (Kitchin 1996). All the preceding is encapsulated in Figs. 8 and 9.

Images Tell a Story

Images such as cyclone forecast track maps in risk communication convey more information than words alone (Goudie 2004; Kitchin 1996). With this feedback from real residents, it becomes clear that linking fire-zone residents, for instance, to educative web fire images becomes an important goal (e.g., BBC 2007).

Community Storytelling Bonds Neighbors

The multipronged approach (Goldammer 2005; Kitchin 1996; King and Goudie 2006) fosters community “storytelling links” (Kikuchi and Nakamori 2007; Lichtenberg and Maclean 1991; Wakefield and Elliot 2003) relating to community self-help and safety. Strengthening community bonds and working with enhanced web resources and community media (Cohen et al. 2006; Quarantelli 2002) will increase community internalization of risk (Finnis et al. 2004), enhancing the likelihood of safety-oriented responses and leading to and possibly including evacuation. People need to accept the reality of the threat, indeed, feel some anxiety about the threat to help drive the intent to seek more information, or the intent to prepare (Paton 2003).

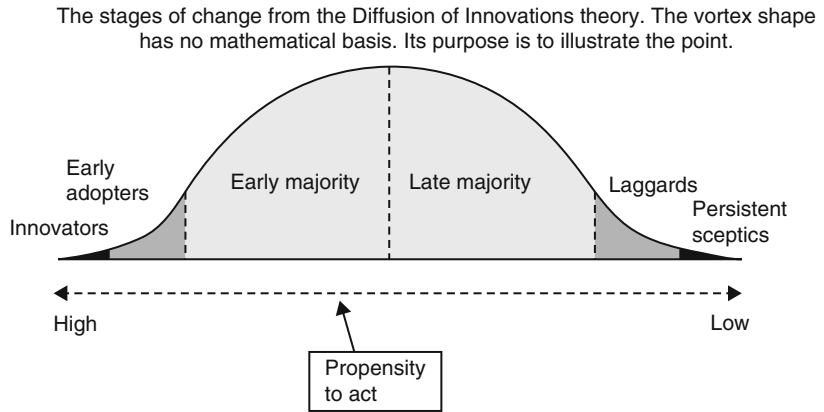
With easier access to relevant web information, with greater detail of current fire behavior and *nowcasts*, fire-zone residents and community radio announcers can describe the looming threat, helping timely preparations and the monumental “stay or go” decision. Community radio, like the national public radio, the ABC, will deliver authoritative and timely risk communication directly from the refined web information.

Theory and emerging practice converge on using refined web-delivered material to households and their neighbors, and to local media, to inform and motivate residents through the whole continuum of the seven steps to community safety. The next section considers institutional barriers to change, followed by extensive research findings and lessons from Australia in disaster management to develop effective warnings and self-evacuations.

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Table 2 Cyclone zone people have much personal contact

Contact with other relatives	% of total response (rounded)
Yes	25
Lots	20
Mobile contact	15
Landline phone	30
No	10
Frequency of neighbor contact	
Often or lots	50
A bit	20
Helped/contacted during eye	15
None	15

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 9 O'Neil's 2004 innovation uptake model



Institutional Barriers to Greater Community Self-Help

Institutional barriers to change take many forms, from “unconsciously” avoiding consideration of the extreme event as “too hard” to a misuse of the power relationships within bureaucracies because of fear of change or malice. Clear examples of the “too hard syndrome” mingled with entrenched vested interests have been denial of links between smoking and cancer or delayed uptake of global warming mitigation.

There were institutional barriers for disaster managers to consider land-based flooding from torrential rains preceding a cyclone (hurricane/typhoon) (Goudie and King 1999) in many cyclone-prone areas. The Queensland State Planning Policy (Phillips 1994) now explicitly refers to probable maximum flood. Including maximum flooding is now part of emergency manager’s internal planning base. This means the exit routes (ER, formula 1) are now considered in planned evacuations.

Since the early 1990s, researchers like Boughton (1992) have suggested having drills for schools and other institutions in readiness for possible earthquakes, cyclones, or other hazards. This author supports precautionary evacuation practices. There are, of course, liability barriers to easily undertaking practice evacuations. There is also uncertainty of threat, which may restrain some risk managers (Svenson 1991).

Institutions have cultures which may passively express antipathy to a paradigm shift toward

sustainability, while being required to usher in sustainability (Dolphin et al. 2000; Goudie 2005). Emergency Management Australia, the peak national body, has the slogan: *Safe, sustainable communities*. However, there is a multitude of conservative forces representing the dominant social paradigm restricting innovation, despite sustainability policy. People engaging in sustainability implementation may come into conflict with institutional representatives of the old paradigm.

new container.

With most development taking place in urban settings, concepts of urban sustainability attempt to merge two different fields of human endeavor – how we modify our landscape, our built environment, and how we behave in that environment. Disaster mitigation is obliged to fuse these two seemingly disparate fields. There may be resistance to that. That resistance needs to be included in any mass disaster movement modeling. The following subsections indicate some reasons for IBs which modelers may need to consider in the IB factor.

Political Insecurity, Real Estate Values, or Undue Alarm

The tension between people’s right to know, government duty of care, and politicians’ perceptions of probability of risk on “their shift” form a complex interplay. If maximizing safety is the shared goal, all risk communication theory suggests people will behave appropriately with the realistic threat information, motivation, and “how to”

instruction on safety-oriented behavior leading into and through and recovering from a natural disaster impact.

Paradigms of Politics: The Real Estate Industry and Vested Interests

Broughton (1992) argues that if people know of the threats, they are likely to support politicians who make sound decisions for community survival. Some plans are made but they are hidden away for various financial and political reasons. This is a case in which attitudinal changes on the part of those communities may change the priorities of the decision-makers and promote the interests of the community.

Those most empowered to assist in implementing projects that require independence, initiative, local links, and knowledge may be the ones who prove most obstructive to innovation (ACF 2004; Jarach 1989; Kobb 2000; Paton et al. 2005).

Action Research (Cronan 1998) sets out to research, develop, and document socially and environmentally desired outcomes within sustainability principles. In this inclusive and “engagement-focused” research, gaining support from the top is crucial to the success of any organizations’ efforts at “social change” (O’Neill 2004). If innovation uptake is viewed as a normative uptake process (Cialdini et al. 1990; Napurrurlarlu and Jakamarrarlu 1988) (Fig. 9), then the gradual, long-term paradigm shift to sustainability (Fien 1993) is plausible and has a conceptual frame.

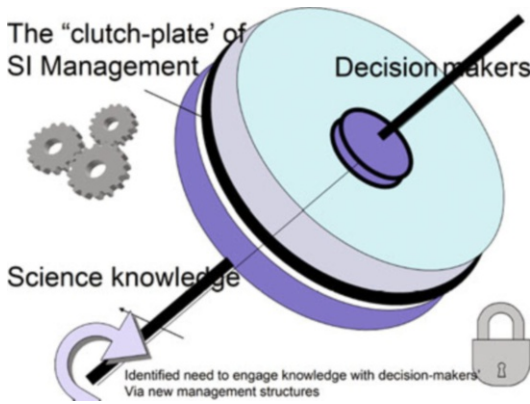
Uptake of the Communication Safety Triangle and the seven steps to community safety (Figs. 7 and 8) will make them the disaster managers’ “norm” over time, simply because that is the direction of social evolution in disaster management. It fits social policy and advanced disaster management approaches, and the technology and willingness of regional media and residents are there. All that is needed is the testing and rollout by more innovative disaster managers. The same applies for Community Safety Groups and variants of *overmanaged, locally controlled near-house social burn-offs* in the case of fire risk management.

Bureaucratic processes need to support the stated goals of their own work units, but individuals may be ambiguous, contradictory, belated, bullying, ill-informed, and, due perhaps to time pressure or arrogance, quite destructive. Such individuals may be “corporate psychopaths” (<http://www.abc.net.au/catalyst/stories/s1360571.htm>). Instead, innovators may be seen as threatening and troublesome. Any innovative pilot project must develop strategies that will maximize change within their unique situation. Ideally, that happens within a supportive parent organization.

In the years since this article was first published, the centrality of shifting social norms, as shown in Fig. 9, has occupied much of my thinking. Institutional barriers to change are critical to many aspects of our collective future. True, there is a propensity for individuals-at-risk not to act – the “it won’t happen to me” syndrome – but there is also this hierarchical disposition to want to hold on to or increase their own power, even at the cost of the very social role their organization purports to serve.

Although organizations I have worked for claim to want to make “evidence-based” decisions, there is a reluctance to do so if such decisions erode their power base. To give some visual depiction of this dilemma, a constructed conflict between knowledge and power, I have developed Fig. 10 as it relates to sustainability implementation in any aspect of our societies. The key issue is the amount of traction between science or evidence-based knowledge and those in senior corporate or government decision-making roles. The model depicted in Fig. 10 argues that the amount of engagement of decision-makers and knowledge gatherers and providers is imperfect.

The value system used here assumes that knowledge should drive structures and behavior to maximize safety and minimize loss to those at risk. When people in bushfire zones, for instance, uniformly say they want real-time, local, and detailed information on which to base the trauma surrounding the decision to evacuate, that is what they want and need (Carlson et al. 2013). Whether decision-makers work to provide that through localized weather stations uploading into the public domain and whether they allow local,



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 10 The imperfect relationship between knowledge and hierarchical decision-makers for sustainability implementation

onground fire chiefs to directly upload local observations and information into the public domain are safety and political issues. Hierarchies want to manage information. Sometimes this is prudent; sometimes this is fatal to those at risk. I only argue that technologies now allow much two-way communication, from reliable and accredited sources, without necessarily channeling and censoring though any “head office.”

The next section discusses the emergent issues of effective risk communication and precautionary self-evacuations.

Experiences and Lessons: Some Case Studies

This section considers disaster impacts and impacts from North Queensland and from fire zones in SE Australia to illustrate the generally positive points of approaches already outlined: living flesh for modelers to understand how subtle and complex but doable successful community preparedness and willingness to act for maximum safety can be. Researching these events, gaining community and disaster managers’ feedback on risk communication, working closely with the Australian Bureau of Meteorology over 15 years, through the federal disaster “paradigm shift” to mitigation (COAG 2004, p. 13), all inform the emergent CST and seven steps to community safety.

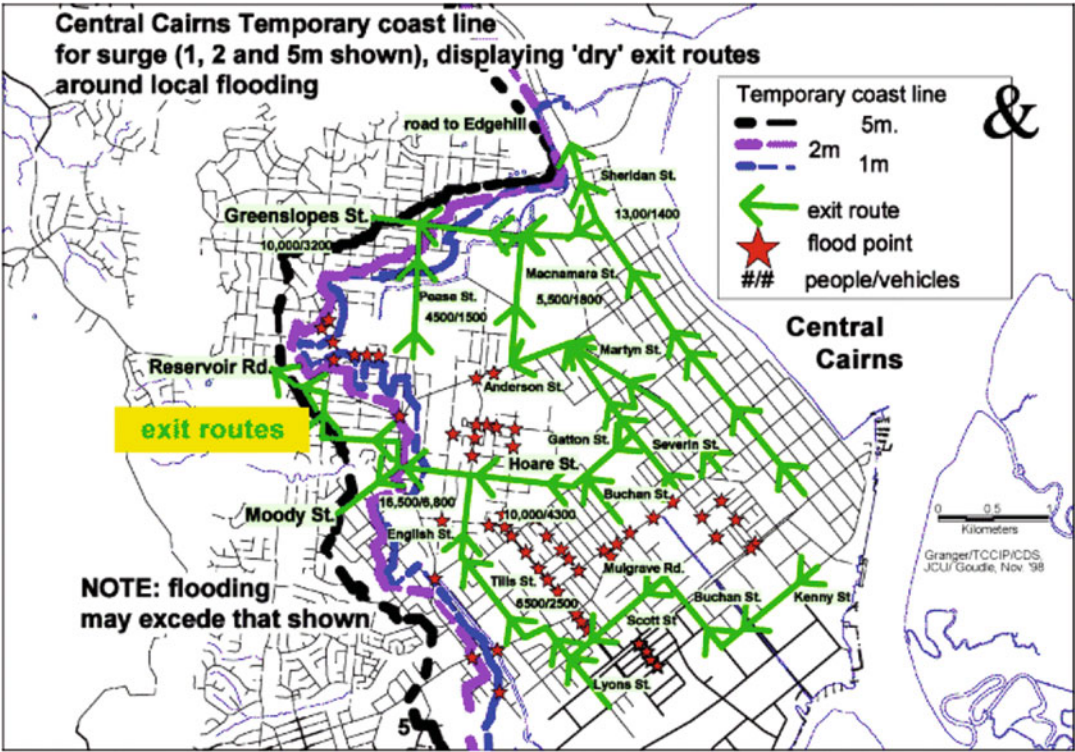
Cairns and Storm Surge Considering Exit Routes (ER, Formula 1)

Cairns City, North Queensland, is centered on land less than 2 m above high tide and subject to cyclone (storm) surge of up to 5 m. A storm surge tracks just behind the eye of a cyclone, a low mound of sea water, perhaps 50 km wide and up to 3 m above normal sea level. It may flow overland for perhaps 4 h, as destructive winds to 280 km/h tear at structures, and churning seas, both laden with pounding debris, behave as battering rams and missiles. Because cyclones have such long lead times from modern electronic detection to landfall, results of this deadly combination (Lichtenberg and Maclean 1991) will be widespread destruction and loss of infrastructure, but not life, with precautionary evacuations (Goudie 2007a; Quarantelli 2002). At the macro-observation scale, people and vehicles in motion within confining environments tend to behave like fluids, capable of modeling. Constrictions of flow paths cause congestion, as described by Helbing et al. (*pedestrian, crowd, and evacuation dynamics*), so early and precautionary evacuation will help minimize the likelihood of gridlock.

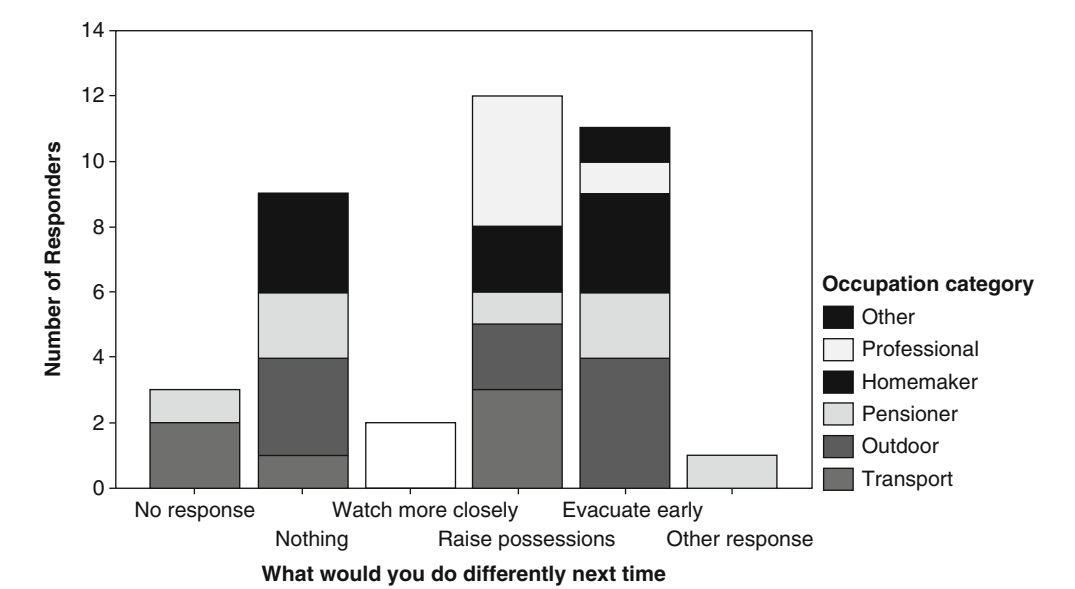
Land-based flooding may be a core issue to effective evacuation (Goudie and King 1999; Goudie 2007a; Yates 1997), with up to 40,000 people in the vulnerable central city area and northern beach suburbs needing early, precautionary evacuation (Salter et al. 1993), Fig. 11. Cairns City Council now has a storm surge map on its public web site (Mornington Peninsula Shire Council 2006), first posted in 2006, like the public and informative flood map for the Redlands Shire, SE of Brisbane, Queensland (Cairns City Council 2006).

Within the philosophical and moral frame of people’s “right to know,” these local governments are using the Internet to inform people they are in a hazard zone, the first step in making the threat real to those residents. They have made the paradigm shift to providing fine-detailed background information which says to residents: “you are in a hazard zone, you may need to do things. Listen for warnings and be prepared to act in a precautionary way.”

Figure 12 shows that after a devastating flood in 1997, the flood-affected residents of Cloncurry,



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 11 Cairns flood-prone evacuation routes and schemata of wave reach and temporary coastline from 1998 information



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 12 Cloncurry reflective of likely responses to future floods

NW Queensland town, would do things differently, with better warnings, when faced with rising flood waters. Remote automatic flood monitoring devices were requested, but the local downpour over a fully flooded, vast, and flat landscape appeared to be the cause of the flood rising 2 m higher than any flood in the prior hundred years. Precautionary evacuation of the low-lying homes would have prevented much heartache and loss of valued possessions (Kitchin 1996).

In 2005, North Queensland was threatened by Cyclone Ingrid. Newspaper-reading residents were left in no doubt about the threat (Fig. 13), a good example of the clear warning role played by the media and nonlanguage image used to convey meaning (section “Discussion”).

A portion of a weather bureau warning media bulletin for Cyclone Ingrid follows, showing,

verbatim, what the nation radio broadcaster, the indigenous radio network, and most responsible media outlet relayed on. This is clear information, including the possible impacts (M and M of formula 1).

Media: For immediate broadcast. Transmitters in the area Cape Grenville to Cooktown are requested to use the Standard Emergency Warning Signal.

TOP PRIORITY

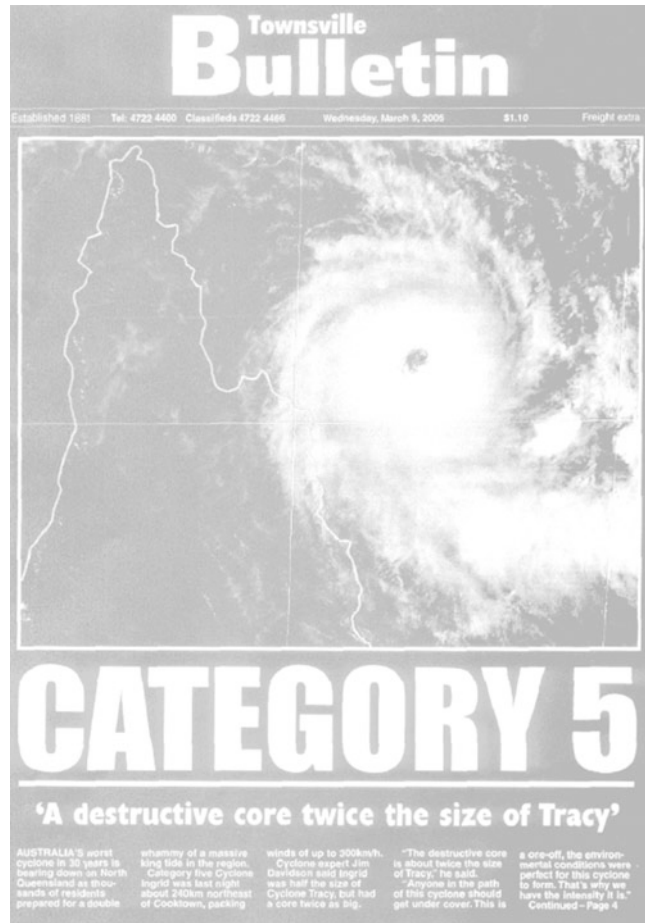
TROPICAL CYCLONE ADVICE NUMBER 14

Issued by the Bureau of Meteorology, Brisbane
Issued at 10:56 am on Wednesday the 9th of March 2005

A Cyclone WARNING is current for communities between Cape Grenville and Cooktown. The warning extends inland across central Cape York Peninsula.

Natural Disasters and Evacuations as a Communication and Social Phenomenon,

Fig. 13 Media coverage of threatening Cyclone Ingrid, 2005



A Cyclone WATCH is current for coastal and island communities on the eastern Gulf of Carpentaria between Weipa and Kowanyama. The watch south to the Gilbert River Mouth has been canceled.

At 10:00 am est Severe Tropical Cyclone Ingrid, category 4, with central pressure 935 hPa, was located near latitude 13.5 south longitude 145.5 east, which is about 140 km northeast of Cape Melville and 260 km east of Coen. The cyclone was moving westward at 11 km/h.

Severe Tropical Cyclone Ingrid poses a serious threat to the far north.

Queensland coast with very destructive wind gusts to 280 km/h near the center. Gales are expected to develop between Cape Grenville and Cooktown during the afternoon. Destructive winds are expected between Coen and Cape Flattery overnight. The very destructive core of the cyclone is expected near the coast between Coen and Cape Melville on Thursday morning.

Coastal residents between Coen and Cape Flattery are specifically warned of the dangerous storm tide as the cyclone crosses the coast early Thursday. The sea is likely to rise steadily to a level significantly above the highest tides of the year with damaging waves, strong currents, and flooding of low-lying areas extending some way inland. People living in areas likely to be affected by this flooding should be prepared to evacuate if advised to do so.

Very heavy rain can be expected to develop on the coast and ranges north of Cooktown.

Research in Port Douglas, NQ, was conducted immediately following Cyclone Ingrid, to help clarify the “boy who cried wolf” hypothesis about “concern fatigue” over precautionary evacuations (section “[Institutional Barriers to Greater Community Self-Help](#)”). Feedback from marine tourist operators and tourist businesses removed about 60 large vessels to safe, up-creek moorings or fully shuttered their premises and raised all stock. These preparations were arduous and costly. The cyclone veered away from Port Douglas, with virtually no impact. Despite that, people were glad of the precautionary evacuation, as a practice (Fig. 14), reinforcing the findings of the Netherlands flood study (Handmer 2000). People

appreciate practice, as part of the new paradigm of a precautionary approach to hazards.

From Those Impacted by Larry, 2006

After Cyclone Larry, a powerful category 4/5 severe cyclone which damaged the Innisfail region, Goudie lead a social science research team into the impact area, interviewing households for 4 days (King and Goudie 2006); from the 150 householders’ survey, focused on risk communication, we learned, as with fire-zone residents and those who experienced Cyclone Larry, that most people, once they have experienced a major disaster, maintain a healthy respect and inclination to act ahead of any future such threat. Hence, the advocated merit of encouraging people with experience to “tell their story”; and community (Media and Message and Community Resilience of formula 1) to help make the threat real to others. The next section is, perhaps, at the leading edge of where disaster management will go: living community self-help. Such communities can provide pilot locations to “calibrate” formula 1.

Woodgate Beach and Community Safety Groups (CSGs)

This section reports a collaborative process between the author and many others, mainly residents of Woodgate Beach, Queensland (Fig. 5), to develop a *preparation and evacuation plan* for residents, formal response groups, the Local Government Disaster Management Group (LDMG), and the Shire Council.

Through the 18-month consultation and development process, the plan needed to be realistic and achievable, relying mainly on locals working together under the LDMG and local SES group, aiming to optimize each element of formula 1.

Community networking and self-responsibility for community safety is profoundly developed in the isolated, 700 strong, central coast Queensland settlement. The volunteer community safety groups were made up of dedicated people. Goudie led the development of a community-based evacuation plan, aiming to ensure maximum preparation and minimum impact on property and residents and optimizing formula 1.

There were five meetings, of up to 80 residents and representatives of all formal response groups, weather and earthquake experts. Meetings included researcher-led 2-day “table-top” evacuation exercise. Woodgate Beach did not develop an evacuation plan but a preparation and evacuation plan

Preparations and Evacuation Self-Help

Approach

To nurture aware, informed residents preparing to be safe through and recover from natural disaster impacts.

Overarching Evacuation Approach

- Identify vulnerable areas or houses. Map those areas.
- Evacuate caravan park, visitors, and people at risk unable to easily move themselves

The Planning Process

1. Define the threats – Fire, flood, wind, cyclone surge, earth movement, and tsunami.
2. Move from where to where? Fires can be fought; floods, storm surge and cyclonic winds, tsunami, or earthquake cannot be. In all threats, make sure

your property is as secure from impact as possible.

3. Identify the vulnerable – Which buildings, infrastructure, and people may be in the threat zone.

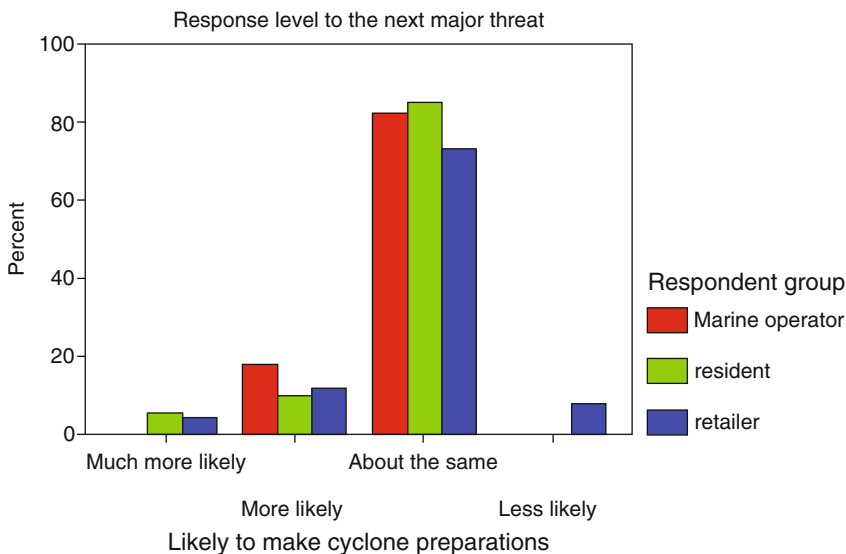
Threats and Treatments

An all-of-community approach has developed an annual round of public education projects, press releases, and pamphlets to inform residents and tourists of the annual cycle of dangers to be wary of, from cyclones to fire care and management. The council newsletter will be a consistent source of information on threats and how to minimize risks (M and M), from the flying debris of a wind storm to being patient at a flood-swollen creek crossing.

A sequenced approach, where the aged and vulnerable are moved first from the highest risk areas, will be taken, as a practice, as a precaution. To highlight the stages of disaster and possible evacuation planning, the background preparation phase is included in this article.

Background Preparations

1. Woodgate Beach residents recognize and act on the need for background preparations to



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 14 People of Port Douglas appreciated a precautionary evacuation

- minimize the impacts of all hazard impacts. This includes property maintenance and upkeep. This maximizes PPR of formula 1.
2. Provide newcomers with an information pack, including a copy of extracts of this community preparations and evacuation plan (CPEP). All community members, including tourists, the elderly, infirm, and needy, are incorporated into this CPEP. This enhances the KB of formula 1.
 3. Provide dot points on evacuation for local residents in the *disaster preparedness information kit*, delivered to each household.
 4. Expose tourists to the essence of local threats and what will be expected of them: leave early, unless they or their vehicles can actively help, under direction.
 5. Define safe shelters – preferably with friends or compatible households. Organizing possible billets for any major impact on portions or all of Woodgate Beach can form a key function of the Community Safety Group. This is the CR of formula 1.
 6. Go through the whole plan, and address matters like the caravan park needing auxiliary power for fuel pumping before the cyclone season.

The Community Safety Group Approach to Disaster Management

This approach is detailed to provide a guide for researchers to use as a framework for other communities.

The Community Safety Group

Purpose Encourage

1. Early Warning Alert

The CSG is an affiliation of existing community groups and neighborhood-level residents who make first contact with “walking-distance” neighbors as soon as anyone hears of a warning that a natural disaster may be approaching their area.

2. Final Preparations (Ramp-Up) Activation

The neighbor-level CSG will provide early local motivation for final safety preparations.

Recovery – Thinking through to recovery (TR) of formula 1.

Recovery is now seen as part of the preparation package, rather than just looking to minimizing impact. The developing approach is to see the whole threat event as one continuous process: from awareness and structural preparedness, through initial communication of threat, to final precautionary preparations and impact and rapid recovery to a fully functional community.

International Snaps

The Center for Disaster Development within the Northumbria University, Newcastle on Tyne, coastal northeast England specializes in recovery and response with an emphasis on development long-term recovery and resilience building. Embedded in this approach is to undertake mitigation. Interviewed by Goudie in May 2005, the director reported: “We use the approach that local knowledge is always drawn on, and that people involved should be in control,” KB and CR of formula 1 and agrees “all disaster management is under the umbrella of sustainable development.” Other interviewed staff support precautionary evacuation as a practice and, independent of the media, an alert signal for people to tune in to the media.

The Shetland Islands

Like Australia’s emerging approach to self-help communities, the Shetland Council’s Emergency Management Planner (2005 interview) gave strong practical support to the approach of an informed, aware, self-activating community ready to act on reliable, clear, how-to-emergency warnings, ranging from making safe if intrinsically out of the main impact zones to early “self-evacuation.” This convergence from differently evolved and slightly different disaster management systems is ground for optimism that the evacuation approach and formula expounded in this article has widely applicable merit.

Remote communities like the Shetlands have lessons for the mainstream in taking responsibility and perforce being self-reliant and oriented to robust self-help. Such isolated communities represent matured examples of the informed, aware communities, predisposed to precautionary action that mainstream populations now

aspire to. Community building is an international aspiration, achievable in urban settings.

Hurricane Katrina

Hurricane Katrina, USA late August 2005 (Fig. 15), has deliberately not been included in this evacuation analysis. With days of clear warning, general formal approaches to precautionary action which underpin the CST and the seven steps to safety and recovery were underplayed.

Figure 15 is included to remind all readers that hazards are real threats to real people in real hazard zones. If all the parameters to maximizing formula 1, such as institutional barriers, are not optimal to safety, great distress and loss can ensue.

This section has provided a cross section of disaster threats, evolving reasons to empower communities, and some factors to include in modeling greater community safety. After discussion in the following section, some recommendations and future directions are provided.

Modeling not only the natural hazard but the various social and communication parameters will provide a good basis for not only simulating impact, say, of the extent of flood waters, but will also highlight which impact areas are the most and least likely to properly act in their own best safety.

Discussion

Risk Communication Theory and Residents in Hazard Zones

People need to know that an impact is possible before they will willingly evacuate. The concept of risk characterization (Handmer 2000; IPA 1997; Paton 2003; Renn and Rohrmann 2000; Rounsefell 1992; Stern 1992) makes clear that people need a practical understanding (*the possibility of impact is real, I fully accept that*) to then illuminate practical choices. “Internalizing” that *a threat is real and what you need to do to maximize your safety* may be the core goal of risk communication. This internalization and safety-oriented action may lead to a choice of precautionary evacuations. People need to make informed decisions, thus leading to decision-driven activity.

Delivering Real-Time Warnings

The idea of subjective uncertainty (Renn and Levine 1991) is well displayed by fire-zone residents who, despite all authority efforts to have them commit to an early decision to stay and defend or to leave their properties early, recurrently reported that they would decide on the day whether they would stay or go. There are tragic and recent international examples of resident’s decisions made on too little understanding or information.

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 15

What no-one should have to stay through – Katrina 2005
http://news.bbc.co.uk/1/hi/in_pictures/4194032.stm



The bushfire “stay or go” decision point explored in 2007 research by the author with Australian fire-zone residents clarifies that decision impact on those under threat – packing up their valuables, children and pets and fleeing their house, almost certainly putting it at risk of burning down, as opposed to staying in their house and experiencing the terror of the developing bushfire – was a decision many preferred to delay. Further, many acknowledged that to leave early with all of the disruption, only to learn that the fire was not an actual threat to their property helped, induces people to delay that decision point.

Subjective uncertainty is a psychological problem. The outcome of current research is to encourage the threat information gatherers and providers, the weather bureau and disaster managers, to refine information detail and use all currently available modes of dissemination to provide the threatened with the maximum amount of fine details to make that evacuation decision in an informed and timely manner. This is a clear example of where theory and the practical views of hazard zone residents converge.

The 2007 southeast Australian fire-zone research shows that people who are more obviously at risk from bushfire are far more prepared than people who are at risk from an occasional but as potentially destructive bushfire. The literature (COAG 2004; ECAP 1997; Finnis et al. 2004) shows that the less frequent the event, the less prepared people are likely to be. Aligned to the goals of this publication, the formula explained in this article on the realities and complex elements of combining physical with human geography to the “social good” of maximizing safety around disaster impacts is a great potential application of complexity and systems science. This complexity includes the media (media support MS of formula 1).

The Media

The media plays an ongoing role in socialization and the development of normative values. Dominick (1993) argues that the media plays a key role in the cognitive development of individuals. Cognition is the act of coming to know something. Mass and social media play defining roles in people’s awareness and responses to disaster threats

(Carlson et al. 2013). The CST embraces local and social media as active agents in providing the needed local information. Local media outlets can draw on and enhance Internet information to inform readers and listeners of “how to” action to maximize their safety. In this way, the mass media can help clarify facts to help make contentious decisions (which may result in life or major property loss) by providing the fine details to reduce the uncertainty surrounding decisions that householders under threat need to make. The seven steps to community safety (section “[Integrating Theory and Implementation](#)”) underline this core need for individuals in hazard zones to accept they are at risk and for the Internet and local media to act as providers of relevant, local, timely information for informed decision-making.

The media can play a powerful role in mobilizing communities (Lichtenberg and Maclean 1991), but they need to have accurate and timely information from disaster managers. The information which fire-zone residents in southeastern Australia in 2007 asked for from the Internet is in full keeping with theory, and a recognition that greater access to current facts will reduce uncertainty in decision-making.

Triggering Action: Modeling and Simulating This Geographic, Communications, and Psychological Complexity

The reason step one in the seven-step process is so important is to counteract a psychological defense against accepting threat. We need to have a sense of personal invulnerability to get us through each day (Renn and Rohrman 2000). If we were scared of all possible problems, we would cease to operate. The risk literature from Handmer, Goudie, Rohrman, Salter (Goudie and King 1999; Handmer 2000; Renn and Rohrman 2000; Rounsefell 1992), and many others underlines the importance of the clarity in the description of the threat and safety-oriented action. Thus, Acceptance they are In a Hazard Zone (AIHZ) of formula 1 is central to possible consequent safety-oriented behavior.

One reason given for risk communication failing is that it is not clear what should be done. The seven steps include a requirement that “what to

do” information be an integral part of any warning as a prelude to evacuation. The clear message is that the process of people excepting that a risk is real needs to well precede any actual impact. This is underlined by Kaspersen and Stallen (1991), who also stress the importance of time perception and time horizons, hence the importance of knowledge base of the hazard and safety-maximizing behavior (KB) of formula 1.

Timely Preparation, Final Actions, and Evacuations

Timeliness of warnings is paramount to the effectiveness of warnings (Svenson 1991). In the early 1990s, there was much psychology research on perceptions of risk probabilities, individual psychometric studies on perceptions of future time, time orientation, planning horizons, and prediction of future events. In 2007 Australian fire-zone residents say they want fine, detailed information as threat approaches, so the uncertainty of what they may face is minimized. Their planning horizons are generally well oriented to maximum safety, but the crucial “stay or go” decision will be based on information as close to impact as allows them adequate time to act safely.

Svenson (1991) and others make clear that it is not only what you may chose to do but also when you do it. In the case of remote area warnings of flooding, for instance, when flood waters may take a week to block down-catchment roads, traveling earlier than planned may be the best way to avoid the fate of the truck shown in Fig. 16. This is an example of an “active warning image,” along with the image of the person standing on the car roof in the flooded road crossing (Fig. 17). Images like these are of use to remind people to travel before or after expected flooding – not during the flood.

In 2007, in a 2-day threat and evacuation exercise in Woodgate Beach operated on the underlined imperative, from the outset, that evacuation ahead of a storm surge would need to be completed not as cyclonic winds struck but 6 h before landfall. Once the winds gusted above 100 km per hour, all peoples, including disaster managers, must be in safe shelter above cyclone surge height.

Disasters and evacuations are issues of people, information, time, and space. People in threat zones are entitled to accurate and timely localized information, detailing the threat and likely time-linked movement. People also want to know, via the Internet or battery-operated radios, what authorities are doing. That information may influence their own actions. What they ask for is the enhancement of their “risk decision landscape” (Sorenson and Mileti 1991).

In 1974 communication management of and responses to disasters in Darwin and Brisbane shocked Australians. Darwin’s Cyclone Tracy killed 65 people, and the Brisbane floods killed 16 (Chamberlain et al. 1981a, b). These catastrophes taught Australian emergency planners much about the importance of effective warnings, sharing honest, complete, and open information in a timely way to emergency managers, emergency workers, and those at risk to ensure sound preparation and responses. The Australian natural disasters of 1974, with major flooding elsewhere in Australia that year, also reinforced the importance of community and family ties to get people through the often profound emotional trauma allied with major natural disasters. The detailed documentation of Cyclone Tracy can help guide development of predictive models of disaster preparedness, communications, and responses. With so few well-documented cases, a Bayesian logic approach could be used, some well-documented disasters, communications, and response cases to develop the model, others to test and refine it.

The more recent and current disaster impacts all point to the importance of communities being properly informed. Indeed, the literature on knowledge, risk, and inclusion of residents in the bushfire planning process is well described by Goldammer (Fowler et al. 2007). Issues of increased encroachment onto bush edges along with climate change are causing increased international concern over fire threats, often to places without much collective wisdom over the reality of those fire threats, or the nature of the strategies needed to maximize community safety.

Southeastern Australia is seen as one of the most bushfire-prone environments in the world (Walmsley 1988). Boura (1998), an Australian



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 16 Nearly crossing a flooded road – Tully NQ 2004, from: Townsville Bulletin, 28/4/4

Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 17 An image dissuading people from driving into flood waters

Conveying weather threat information - flood



fire manager, describes the development of community fireguard groups in Victoria. These groups are managed and instructed by the Country Fire Authority, but they are residents within very localized neighborhoods. They have their counterparts in the ACT and some other Australian states. They draw on detailed, often technically advanced, information from the weather bureau and fire agencies to increase local awareness and action. This is a real playing out of the theories of social amplification developed by psychologists and risk communication theorists (Raggatt et al. 1993), advanced planning techniques (Cuthill 2004) and aligned with the way many feel about the direction of emergency management in Australia (Cronan 1998).

Linking the People with the Message

The foci on culture, community, and social frameworks, either formal or informal neighborly links, are considered by Douglas (1992). Douglas argues that there is a need for people to

understand that risk and danger exist where they live, despite a low probability that an impact may occur in any given hazard season. There is a large body of psychology which considers why people ignore clear messages which may maximize their safety (Cialdini et al. 1990). Douglas speaks of artificially distorted world views. Douglas posits such bias is rooted in oversimple views of heroic and bourgeois fiction. Changing such a normative world view has been discussed in this article in terms of a paradigm shift, well displayed by the Australian Government (COAG 2004). A model constructed from formula 1 needs to include all the subtle complex issues of social profiling.

Woodgate Beach (section “Experiences and Lessons: Some Case Studies”), perhaps because of physical isolation, is fully embracing self-help and also displaying a deep culture of volunteerism that can be nurtured and emulated in “urban villages.” Sustainable urban planning concepts of nodes or activity centers are now well evolved as

the hub of “urban villages” (Goudie 2001). What is happening in Woodgate Beach can be used as a model for any formula of developed social collective will to self-help, ultimately whether urban and surrounded by other neighborhoods or more physically isolated.

Like the Ferny Creek (Victoria) community who agitated for a fixed bushfire siren after three of their neighbors were burned to death in the flash fire of 1997, community action needs a few individuals of vision and drive; residents of “neighborhoods” can initiate and embrace safety-oriented behaviors and structures. The best thing the institutional systems can do is draw out, nurture, and encourage these individuals to mobilize a focus on gaining threat knowledge and acting to increase community responses. Part of the community engagement in the fact they are in a threat zone, as step 1 to action, is shown in Fig. 18, a great roadside banners initiative by one section of the Victorian Country Fire Authority to make the fire threat and implications real for fire-prone residents. Figure 18 shows, in few words, that if you intend to evacuate, do it early. Figure 18 says that the support and creativity to maximize communication effectiveness (the opposite of Institutional Barriers) help contribute to a positive Medium and Message (MM) of formula 1.

Traditional Weather Warnings

Traditional weather warning signs used by Australian Aborigines include ant movement for looming flood and general bird movement for strong winds, including cyclones (Munro 1995; Svenson 1991; Thompson 2002). On Palm Island (Fig. 5), when the birds and animals go quiet, it signals that a major storm may be on the way (Goudie 2004; Utemorrhah 1980). The buildup to the summer monsoon rains is the universal experience of still, humid and hot conditions, continuing in intensity and cloud until the rains come. All cultures in all hazard zones have their traditional knowledge. Cultural knowledge bases (KB of formula 1) needs to be modeled into any mathematical prediction of likely propensity to respond to a natural disaster threat.

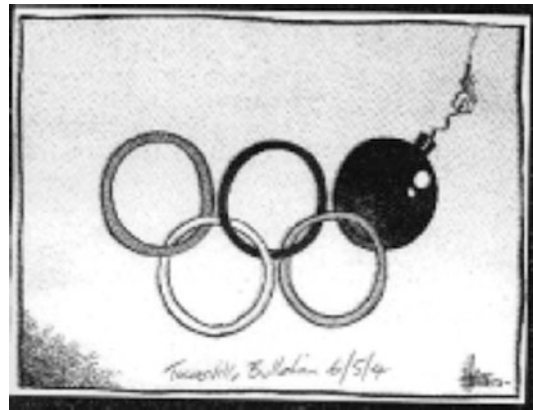
Words and Images

Word and image use are critical to effective communication. A newspaper graphic ahead of the terrorist-threatened 2004 Olympic Games (Fig. 19) conveys much about the world we now occupy and uses no words.

The need for clear messages, most likely to provoke a precautionary response, is supported by the risk communication literature of section “Effective Risk Communication” and the communication and cognitive theory of section “Integrating Theory and Implementation.” If this knowledge is melded to requests of remote



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 18 Initiatives to empower residents to be prepared for fire



Natural Disasters and Evacuations as a Communication and Social Phenomenon, Fig. 19 An image of terror in 2004 – Source: Townsville Bulletin

indigenous peoples and NESH, the safety goal of clear, plain words and images will become the risk communication norm. There are many examples of images conveying meaning, such as the internationally used figure running upstairs, with an arrow, signifying an evacuation route. The symbol for tsunami, a stylized figure racing up a steep slope with a very large wave following, threatening to engulf them, is also without words but conveys all we need to know about evacuating ahead of a tsunami: get up the slope immediately.

Images and Warning Maps

Risk communication is usually about attempts to prompt considered action by a person or community. Effective communication should make the future threat real in present thinking. Alternative responses should be outlined, along with likely consequences. To further prompt a considered and active response to the “action warning,” the consequences of inaction or a range of defensive actions should be made lucid. “Preferred” behavior should seem reasonably attractive to “target individuals” (Sullivan 2003). The theoretical overview of risk communication in section “Effective Risk Communication” explains why the message must be clear to the target audience. It needs to have some cognitive content to get people thinking about how it may impact on them and what the alternative outcomes for them may be if the predicted impact strikes where they are. Stimulus and local risk simulations, as embedded in the Redlands shire flood map (Cairns City Council 2006), should be saturation broadcast into all hazard zones. The CST should be attempted with at-risk groups, people in threat zones, encouraging them to properly think about the real threat and to indicate to people what inaction may bring. A range of safety-oriented actions should also be presented.

Modeling Risk Communication into and Through Communities

Efforts have been made since the 1960s to see how well people understood the hazard. These studies continue (Berry and King 1998). The model of hearing, understanding, believing, and feeling that the information is personalized so that

those at risk will act has been well understood since the early 1990s (Sorenson and Mileti 1991). The general issues of credibility (DIMIA 2004; Paton et al. 2005; Phillips 1994) still apply. Studies reported by Sorenson and Mileti (1991) show increased knowledge as a result of risk communication efforts. People do become more aware of hazards and their personal place within the hazard threats. Unfortunately, the link between knowledge and behavior remains tenuous (Goudie 2001).

Knowing that the normal first warning of major disruptive weather comes via the evening TV news in remote settlements, simulations will carry a high embedded likelihood of safety-oriented community response.

Sorenson and Mileti (1991) showed that the believability of warnings increases as people get more warnings from officials with high credibility and that women tend to believe emergency warnings more than men. They also showed that people higher up the socioeconomic ladder tend to believe warnings more than their counterparts. Minority groups have lower than average belief in warnings, while people with a high knowledge of the hazard tend to find warnings more credible.

New Cultures to Hazard Zones

The issues discussed in this article from remote indigenous communities and from non-English-speaking households make clear the importance of using plain English. Work with representatives of both groups made clear that plain English and images were necessary to convey the concept of danger to people in hazard zones, concurring with the expositions of Douglas (1992). The simpler the language, the better. Douglas argues that even the word “risk” could be dropped and “danger” used instead. Goudie’s work with recent Somali arrivals in Australia showed that, like Japanese tourists and many other people from non-English-speaking backgrounds, they understood the word and concept of “danger,” but did not, for instance, understand the word “severe.” The importance of language clarity *to the intended audiences* in the MM expression of formula 1 needs emphasis.

For risk communication and evacuation advice to penetrate to all people in hazard zones, the language and images must be clear, simple, and compelling. If risk communicators construct campaigns of engagement which work for such “marginalized” groups, the “mainstream” can be easily included in that campaign.

Playing the Odds

Renn and Rohrmann (2000) make clear that the assumption that risk judgments and evaluations are universal processes independent of social status or national heritage is unreasonable. Goudie’s Australian research with multicultural organizations shows most clearly that there are many populations in hazard zones and that each of those groups must be addressed in ways that make the hazard real to them. Findings from the Cyclone Larry research (King and Goudie 1998, 2006) also show that there are many groups other than cultural, such as “social isolates” – the socially disenfranchised, who need special attention around disaster threats to ensure their safety.

Douglas (1992) argues that risk perception and thus response are also issues of moral and political issues. Westerners tend to consider the probability of an impact in terms of gambling, whereas other cultures may need other triggers to internalize the threat and motivate them to action. The concept of a low probability event needing to be taken seriously is discussed by Douglas and others. Residents of Florida are well used to a near-annual evacuation ahead of frequent cyclones. Residents of the coastal communities of Queensland may be less likely to take the warning seriously, because they believe that there is no real chance that they will be personally affected by a cyclone. The whole issue of low probability and high impact events needs to be stressed to relevant impact zone residents.

The first and hardest step in an effective community evacuation remains convincing people in the projected impact zone that the looming threat is real. Renn and Rohrmann (Renn and Levine 1991) suggest that to help make that threat real, project the expected number of fatalities or the catastrophic potential or where the threat may come from.

Spreading the Warning

Handmer (2000) recommends that the professional warning agencies should attempt to harness the “personal informational networks of individuals within formal communication systems, and by assuring that formal warning advice is consistent with local norms and behavior” (Handmer 2000, p. 27). Shifting the normative values of recalcitrant disaster managers and residents in ignorance or denial in urban risk zones becomes the key task of the *seven steps to community safety*, along with the amalgamating approach of the *community safety triangle*.

Part of the strength of the CST is that it taps into and informs an existing social predisposition for people to talk to each other, particularly in times of common threat (King and Goudie 2006). Using emergent technologies to provide real-time information to community media and households helps realize aware, informed communities, predisposed to action.

With literature and research results discussed throughout this article, this discussion underlines the importance of modeling local norms and interactive, safety-focused behavior. Emerging concepts which may be tested in further research are community safety groups (section “[Experiences and Lessons: Some Case Studies](#)”) and social burn-offs or flood evacuation practices. The main thrust of this work is to encourage interested modelers to accurately simulate how communities understand and act on the need for safety-oriented action, where people are inclusive at the neighborhood level, acting as a self-preserving group, no matter what the ethnicity or state of surrounding populations. Neighborhood cohesion and empowerment is the current social policy consistent with ESD philosophies and principles, enshrined in current planning laws such as IPA (1997).

To illustrate the power of the Internet, consider that any online reader globally can click on ([Redlands Shire Flood Maps](#)) and see the social aspects of Queensland’s sustainability planning law. Indeed, click on and draw from any of the web sites given. If we were not suffering from information overload and time paucity, the paradigm shift to sustainability implementation would be rapid, the modeling acted on by disaster

managers and financial cost-benefit analysts. Until that time, models will help convince and guide people in testable and demonstrable ways to increase community safety. Our very busyness is perhaps the main barrier to reducing global warming and natural disaster impacts. An alternative way into a more sustainable future is centered on local needs meeting, including nurturing more local community cohesion and sharing of reliable, safety-oriented information and safety-oriented actions.

Future Directions in Modeling Disaster Preparedness and Evacuations as a Social Issue

As global warming and climate change intensifies, the need to model for preparedness and evacuations will increase with more frequent and extreme weather events. The total and fine detail of all needed background preparations and ramp-up preparations are too numerous and arduous for formal disaster management organizations to implement alone, hence the increased need to promote and strongly support the role of community engagement, of community empowerment and nurturing self-help in maximizing effective natural disaster preparation, including evacuations.

Australia has matured disaster policy, law, and evolving practice, all embedded in concepts of ecologically sustainable development. Researchers can access, model, and test local applicability of some of the Australian experiences and culture of community self-help. Sustainability implementation research will become universal as the era of last-minute organizational flurries, while goading an ignorant and passive population at threat is superseded by prepared and bonded communities who are primed to receive well-delivered warnings, often sourced from the web, and who move themselves to safety in plenty of time. Modeling this will help sell the approach to conservative disaster managers.

The sustainability implementation research, including data-based simulations introduced in this article, is the logical next step to “action research” of the social sciences and will become the norm in approaching all issues of sustainability where the policies are mature, but agencies are

unsure how to implement the paradigm shift, the behavioral and technological shift, to agreed long-term goals.

Meaningful consultation is a defining requirement of sustainability planning and good modeling, so all future social research of merit will take the human geographer’s approach and “ask the local residents, involve the local residents.” Community empowerment, and its modeling, means local residents are entitled to help mold their own future. With hazard management, the shared SIR goal is to empower hazard zone communities to be aware of the threat, be basically prepared for any warning, and act, as a community, in a precautionary way to maximize safety, minimize loss, and speed recovery.

As we move forward, plain language(s) and clear and widely displayed hazard maps and images of like hazard impacts will help energize hazard zone residents to the threats and their own needed safety-oriented behavior.

Researchers and complexity and systems scientists have a moral, if onerous, obligation to challenge any layer of government clearly exercising or imposing barriers to helping people internalize any threats and then be supported in getting safe and staying safe, recognizing that change may threaten some individuals or sections within bureaucracies. As disaster management moves from a militaristic model of command and control to community empowerment and self-help, the developing potential of the web as a key information source into hazard zones – ultimately web-to-air – will bear the fruit of greater community safety and minimized loss. Web-to-TV in disaster warning is a fine conduit for showing residents-at-risk where the threat may be in relation to their home in a few hours. That immediacy of moving images is a compelling motivator for safety-oriented action.

In the future, researchers, modelers, and others involved in maximizing community safety will embrace some variations of the *communications safety triangle* and the *seven steps to community safety*, simply because they make sense and are highly cost-effective and easily web-refined from international to local conditions, populations, and threats.

Acknowledgments I most thank my research guardian and mentor over 15 years, Prof. David King, director of both the Australian Center for Disaster Studies and of the Center for Tropical Urban and Regional Planning. David allowed me freedom to develop as an “evidence-based” scientist, to conceive core approaches to sustainability implementation research, productive in helping render positive change in both disaster management and sustainability planning. Thanks to Australian Bureau of Meteorology staff for their interactive support to improve risk communications, listening to what real people in real hazard zones experience, how they hear warnings, and how the medium and the messages can be and are optimized. The Bureau embraces the core goal to motivate safety-oriented action by people in hazard zones. The Bureau listens and improves the message and the delivery. The bushfire research of 2006 and 2007 was funded by the Australian Bushfire Cooperative Research Center, supported by the University of Tasmania. The 14 years of research reported in this article was not possible without the contributions of authorities and more than 1000 Australians, old and new, who opened their organizations or doors to myself or research team members and shared their hazard experiences and specific warning needs. Thanks all.

Bibliography

Primary Literature

- ACF (2004) Institutions for sustainability. Australian Conservation Foundation ACF, pp 1–37
- AMCORD (1995) AMCORD95, Australian model code of residential development. Department of Housing and Regional Development. Aust Govt Printing Service, Canberra
- Australian Bushfire CRC (2005) Bushfire CRC/research/community self sufficiency for fire safety/effective risk communication. <http://www.bushfirecrc.com/research/c41/c41.html>. Accessed 2008
- Baram M (1991) Rights and duties concerning the availability of environmental risk information to the public. In: Kasperson RE, Stallen PJM (eds) Communicating risks to the public – international perspectives. Kluwer, Dordrecht, p 481
- BBC (2005) <http://news.bbc.co.uk/1/hi/world/europe/6963373.stm>. Access to 2008, In pictures: Hurricane onslaught
- BBC (2007) <http://news.bbc.co.uk/2/hi/americas/7055721.stm>. Californians flee as fires rage
- Berry L, King D (1998) Tropical cyclone awareness and education issues for far North Queensland school students – storm watchers. Aust J Emerg Manag 13:6
- Boughton GN (1992) Education on natural hazards, the Macedon digest. Aust J Disaster Manag 7(2):4–7
- Boura J (1998) Community fireguard: creating partnerships with the community to minimise the impact of bushfire. Aust J Emerg Manag 13:59–64
- Bruntland GH (1988) Our common future. The World Commission for the Environment and Development. Alianza Publications. http://tilastokeskus.fi/abo2004/foredrag/hoglund_pp.pdf, <http://web.uvic.ca/~stucraw/Lethbridge/MyArticles/Brundtland.htm>. Accessed 2008
- Cairns City Council (2006) Storm tide maps. <http://www.hazardsaustralia.info/Mapping.html>
- Cairns City Council (2007) Storm surge maps. http://www.cairns.qld.gov.au/_data/assets/pdf_file/0012/20019/cairns-cbd-barron-trinity-inlet-1-301.pdf Accessed 2013
- Carlson JM, Alderson DL, Stromberg SP, Bassett DS, Craparo EM, Gutierrez-Villarreal F, Otani T (2013) Measuring and modeling behavioral decision dynamics in collective evacuation. *arXiv preprint arXiv:1304.4704*
- Chamberlain ER, Hartshorn AE, Muggleston H, Short P, Svensson H, Western JS (1981a) Queensland flood report Australia day (1974). Australian Government Publishing Service, Canberra, p 38
- Chamberlain ER, Doube L, Milne G, Rolls M, Western JS (1981b) The experience of cyclone Tracy. Australian Government Publishing Service, Canberra, p 150
- Cialdini R, Reno R, Kallgren C (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J Pers Soc Psychol 58(6):1015–1026
- COAG (2004) Natural disasters in Australia, reforming mitigation, relief and recovery arrangements. [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(756EDFD270AD704EF00C15CF396D6111\)-COAG+Report+on+Natural+Disasters+in+Australia+-+August+2002.pdf/\\$file/COAG+Report+on+Natural+Disasters+in+Australia+-+August+2002.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(756EDFD270AD704EF00C15CF396D6111)-COAG+Report+on+Natural+Disasters+in+Australia+-+August+2002.pdf/$file/COAG+Report+on+Natural+Disasters+in+Australia+-+August+2002.pdf). Accessed 2008, Council of Australian Governments Commonwealth of Australia, 201
- Cohen E, White P, Hughes P (2006) Bushfire and the media reports 1–3. Latrobe University and BCRC, La Trobe University
- Cronan K (1998) Foundations of emergency management. Aust J Emerg Manag 1(13):20–23
- Cuthill M (2004) Community well-being – the ultimate goal of democratic governance. Qld Plan 44(2):8–11
- Cutter SL, Boruff BJ, Shirley WL (2003) Social vulnerability to environmental hazards. Soc Sci Q 84(2): 242–261
- DIMIA (2004) Department of Immigration, Multicultural and Indigenous Affairs
- Dolphin RR, Richard R, Ying F (2000) Is corporate communications a strategic function. Manag Decis 38(2):99–106
- Dominick JR (1993) The dynamics of mass communication. Mc Graw-Hill, Columbus, p 616
- Douglas M (1992) Risk and blame. Essays in cultural theory. Routledge, London
- Drabek TE (1994) Disaster evacuation and the tourist industry, Program on environment and behaviour monograph 57. University of Colorado, Colorado
- Dunlap RE, Van Liere KD (1978) The ‘new environmental paradigm’: a proposed measuring instrument and preliminary results. J Environ Ed 9(4):10–19
- ECAP (1997) Guidelines for disaster prevention and preparedness in tropical cyclone areas. Economic Commission for Asia and the Pacific, the World

- Meteorological Organisation and the Red Cross Societies, Geneva
- EMA (2000) Emergency risk management applications guide, The Australian Emergency Manuals Series. Emergency Management Australia, Dickson, p 4 Emergency Management Australia
- EMA (2002a) Research agenda for emergency management. March. EMA research and development strategy for "safer sustainable communities". Emergency Management Australia, Dickson
- EMA (2002b) Indigenous communities and emergency management. Emergency Management Australia, Canberra, p 22
- EMA (2002c) Guidelines for emergency managers working with culturally and linguistically diverse communities. <http://www.em.gov.au/Documents/Manual44-GuidelinesforEmergencyManagementinCulturallyandLinguisticallyDiverseCommunities.pdf>. Accessed 2013
- EMA (1998) Emergency management Australia information service. Report of the strategic planning conference on the development of enhanced awareness education programs and materials for remote Aboriginal and Torres Strait Islander communities. Darwin, May 1997. In: Conference proceedings, EMA, Mt Macedon
- Fien J (1993) Education for the environment – critical curriculum theorising and environmental education. Deakin University, Geelong
- Finnis K, Johnston D, Paton D (2004) Volcanic hazard risk perceptions in New Zealand. Tephra, earth and atmospheric sciences. University of Alberta, Edmonton, pp 60–64. [http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/\\$file/volcanichazardrisk.pdf](http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/$file/volcanichazardrisk.pdf), 2008
- Fowler KL, Kling ND, Larson MD (2007) Eigenvalues. Bus Soc 46:1. March. 88–103. Sage Publications
- Goldammer J (2005) Wildland fire – rising threats and vulnerabilities at the interface with society. In: United Nations Tudor Rose, Jeggle T, United Nations (eds) Know risk. UN International Strategy for Disaster Reduction (UN-ISDR), Geneva, pp 322–323/376 p
- Goudie D (2001) Toward sustainable urban travel. PhD thesis, James Cook University. <http://eprints.jcu.edu.au/967/>, 2008
- Goudie D (2004) Disruptive weather warnings and weather knowledge in remote Australian indigenous communities. Web-based report. http://www.jcu.edu.au/cds/public/groups/everyone/documents/technical_report/jcutst_056224.pdf, 2008
- Goudie D (2005) Sustainability planning: pushing against institutional barriers. Ecosystems and sustainable development V. WIT Press. WIT Trans Ecol Environ 81(5):215–224. www.witpress.com, 2008
- Goudie D (2007a) Transport and evacuation planning. In: King D, Cottrell A (eds) Communities living with hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, Townsville, pp 48–62. 293
- Goudie D (2007b) Oral histories about weather hazards in northern Australia. In: King D, Cottrell A (eds) Communities living with hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, Townsville, pp 102–125. 293
- Goudie D, King D (1999) Cyclone surge and community preparedness. Aust J Emerg Manag 13(1):454–460. http://www.em.gov.au/Documents/Cyclone_surge_and_community_preparedness.pdf
- Gurmankin AD, Baron J, Armstrong K (2004) Intended message versus received message in hypothetical physician risk communication: exploring the gap. Risk Anal 24(5):1337–1347
- Handmer J (1992) Can we have too much warning time? A study of Rockhampton, Australia. The Macedon digest. Aust J Disaster Manag 7:2. p 8–10
- Handmer J (2000) Are flood warnings futile? Risk Commun Emerg 2(e):1–14. <http://www.massey.ac.nz/~trauma/issues/2000-2/handmer.htm>, 2008
- Handmer J (2001) Improving flood warnings in Europe: a research and policy agenda. Environ Hazards 3(2001): 19–28
- Heller K, Alexander DB, Gatz M, Knight BG, Rose T (2005) Social and personal factors as predictors of earthquake preparation: the role of support provision, network discussion, negative affect, age, and education. J Appl Soc Psychol 35(2):399–422
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14(4):382–417
- IPA (1997) The integrated planning act. Queensland Government, Now Sustainable Planning Act 2009. <http://www.legislation.qld.gov.au/legisltm/acts/2009/09ac036.pdf>, 2013
- ISR (2007) Australian policy online. Institute for Social Research, Swinburne University of Technology, [1] <http://apo.org.au/>
- Jarach M (1989) Overview of the literature on barriers to the diffusion of renewable energy sources in agriculture. Appl En 32(2):117–131
- Kasperson RE, Stallen PJM (1991) Communicating risks to the public international perspectives. Kluwer, Dordrecht
- Kikuchi T, Nakamori Y (2007) Agent model analysis to explore effects of interaction and environment on individual performance. J Syst Sci Complex 20:1–17. Springer Science + Business Media, Inc
- Kim YC, Ball-Rokeach SJ (2006) Civic engagement from a communication infrastructure perspective. Commun Theory 16:173–197
- King D, Goudie D (1998) Breaking through the disbelief – the March 1997 floods at Cloncurry. Even the duck swam away. Aust J Emerg Manag 4(12):29–33
- King D, Goudie D (2006) Cyclone Larry, March 2006 post disaster residents survey. Centre for Disaster Studies, James Cook University, with the Australian Bureau of Meteorology P77. http://www.jcu.edu.au/cds/public/groups/everyone/documents/technical_report/jcutst_056193.pdf
- King D, Goudie D, Dominey-Howes D (2006) Cyclone knowledge and household preparation – some insights

- from cyclone Larry report on how well Innisfail prepared for cyclone Larry. *Aust J Emerg Manag* 21:3, 52–59. <http://www.em.gov.au/Documents/Cyclone%20knowledge%20and%20household%20preparation.pdf> Accessed June 2013
- Kitchin RM (1996) Increasing the integrity of cognitive mapping research: appraising conceptual schemata of environment-behaviour interaction. *Prog Hum Geogr* 20(1):56–84
- Kobb P (2000) Emergency risk management applications guide. Emerg Manag Aust, Dickson: Emergency Management Australia. Australian Emergency Manuals Series; 05. <http://www.em.gov.au/Documents/Manual%2005-ApplicationsGuide.pdf>
- Leibovitz J (2003) Institutional barriers to associative city-region governance: the politics of institution-building and economic governance in 'Canada's Technology Triangle'. *Urban Stud* 40(13):2613–2642
- Lewis C (2006) Risk management and prevention strategies. *Aust J Emerg Manag* 21(3):47–51
- Lichtenberg J, Maclean D (1991) The role of the media in risk communication. In: Kasperson RE, Stallen PJM (eds) Communicating risks to the public – international perspectives. Kluwer, Dordrecht, p 481
- Lidstone J (2006) Blazer to the rescue! The role of puppetry in enhancing fire prevention and preparedness for young children. *Aust J Emerg Manag* 21(2):17–28
- Loudness RS (1977) Tropical cyclones in the Australian region July 1909 to June 1975. AGPS, Canberra
- McKenna F (1993) It won't happen to me: unrealistic optimism or illusion of control. *Br J Psychol* 84:39–50
- Mornington Peninsula Shire Council (2006) Fire wise fire management. http://www.mornpen.vic.gov.au/Services_For_You/Fire_Emergency_Management/Fire_Management_Plans
- Munro DA (1995) Ecologically sustainable development – is it possible? How will we recognise it? In: Sivakumar M, Messer J (eds) Protecting the future – ESD in action. Futureworld, Wollongong
- Napurrurlarl NO, Jakamarrarl NP (1988) Ngawarra-Kurlu. Yuendumu B.R.D.U., Darwin, p 19
- NDSI Working Group (2000) Effective disaster warnings. Working Group on Natural Disaster Information Systems. Subcommittee on Natural Disaster Reduction National Science and Technology Council Committee on Environment and Natural Resources. Executive Office of the President of the United States of America, 56
- O'Neill P (1999) Social marketing. <http://www.ses.nsw.gov.au/resources/research-papers/communities/a-social-marketing-framework-for-development-of-public-awareness>. NSW Govt. Australia
- O'Neill P (2004) Why don't they listen – developing a risk communication model to promote community safety behaviour. The International Emergency Management Society, 11th Annual Conference Proceedings, Melbourne, May 18–21 2004. pp 160–169
- Paton D (2003) Stress in disaster response: a risk management approach. *Disaster Prev Manag* 12(3):203–209
- Paton D, Smith L, Johnston D (2005) When good intentions turn bad: promoting natural hazard preparedness. *Aust J Emerg Manag* 20:25–30
- Phillips R (1994) Long range planning. *London* 27(4): 143–145
- QG & QES (2003) State planning policy. Mitigating the adverse impacts of flood, bushfire and landslide. State planning policy 1/03. Dept Local Government and planning, & Dept of Emergency Services. <http://www.emergency.qld.gov.au/publications/spp/pdf/spp.pdf>
- Quarantelli EL (2002) The role of the mass communication system in natural and technological disasters and possible extrapolation to terrorism situations. *Risk Manag Int J* 4(4):7–21
- Queensland Government (1997) Integrated planning act. <http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/S/SustPlanA09.pdf>
- Raggatt P, Butterworth E, Morrissey S (1993) Issues in natural disaster management: community response to the threat of tropical cyclones in Australia. *Disaster Prev Manag* 2(3):12–21
- Redlands Shire Flood Maps. <http://maps.redland.qld.gov.au/website/redemapexternal%5Fv2%5F03/Default.aspx>
- Renn O, Levine D (1991) Credibility and trust in risk communication. In: Kasperson RE, Stallen PJM (eds) Communicating risks to the public – international perspectives. Kluwer, Dordrecht, p 481
- Renn O, Rohrmann B (2000) Cross-cultural risk perception, a survey of empirical studies. Kluwer, Dordrecht, p 240
- Rohrmann B (2000) A socio-psychological model for analysing risk communication processes. *Aust J Disaster Trauma Stud* 2000(2). <http://www.massey.ac.nz/~trauma/issues/2000-2/rohrmann.htm>
- Rounsefell V (1992) Unified human settlement ecology. In: Birkeland J (ed) Design for sustainability: a sourcebook of integrated eco-logical solutions, vol S4.2. Earthscan Publications, London, pp 78–83
- Salter J (1992) The nature of the disaster – more than just the meanings of words: some reflections on definitions, doctrine and concepts. The Macedon digest. *Aust J Disaster Manag* 7(2):1–3
- Salter J, Bally J, Elliott J, Packham D (1993) Natural disasters: protecting vulnerable communities. In: Merriman PA, Browitt CW (eds) Conference proceedings. Royal Society (Great Britain), London 13–15 October 1993
- Sheppard E (1986) Modelling and predicting aggregate flows. In: Hanson S (ed) The geography of urban transportation. Guilford Press, New York, pp 91–110
- Skertchly A, Skertchly K (2000) Message sticks – hazard mitigation visual language. EMA, ACT, Dickson, Australian Capital Territory
- Sorenson J, Mileti D (1991) Risk communication in emergencies. In: Kasperson RE, Stallen PJM (eds) Communicating risks to the public – international perspectives. Kluwer, Dordrecht, p 481
- Stern P (1992) Psychological dimensions of global environmental change. *Annu Rev Psychol* 43:269–302

- Stern PC, Fineberg HV (1996) Understanding risk, informing decisions in a democratic society. National Academy Press, Washington, DC, p 249
- Sullivan M (2003) Communities and their experience of emergencies. *Aust J Emerg Manag* 18(1):19–26
- Svenson O (1991) The time dimension in perception and communication of risk. In: Kasperson RE, Stallen PJM (eds) *Communicating risks to the public – international perspectives*. Kluwer, Dordrecht, p 481
- Thompson KM (2002) Variability and uncertainty meet risk management and risk communication. *Risk Anal* 22(3):647–654
- Utemorrhah D (1980) How the people were all drown. In: Mowanjum (ed) *Visions of Mowanjum: aboriginal writings from the Kimberley*. Rigby, Adelaide
- Utemorrhah D, Clendon M (2000) Dumbi the owl. In: Kimberley Language Resource Centre (ed) *Worrorra Lalai, Worrorra dreamtime stories*. KLRC, Halls Creek, p 113
- Wakefield SE, Elliot SJ (2003) Constructing the news: the role of local newspapers in environmental risk communication. *Prof Geogr* 55(2):216–266
- Wall M (2006) The case study method and management learning: making the most of a strong story telling tradition in emergency services management education. *Aust J Emerg Manag* 21(2):11–16
- Walmsley DJ (1988) *Urban living, the individual in the city*. Longman scientific and technical, Longman, London, p 104
- Woods F, Gabriel P (2005) Individual responsibility and state-wide strategies: bushfire in Victoria, Australia. In: Jeggle T (ed) *Know risk*. United Nations, Tudor Rose, pp 326–328. Leicester, LE1 5RA, UK 376
- Yates J (1992) Assisting the community to plan: a pilot program in Western Australia. *The Macedon digest*. *Aust J Disaster Manag* 7(2):12
- Yates J (1997) Federalism and disaster mitigation in remote aboriginal communities in Western Australia. *Spring, AJEM*, 25–32, Emergency Management Australia, Mt Macedon Victoria Australia. Publisher: Grey World-wide Canberra Australia
- Yeo S (2002) Natural hazards. Flooding in Australia: a review of events in 1998. 25:177–191. Department of Physical Geography, Macquarie University, NSW, <http://www.springerlink.com/content/n160g0121800n742/> Natural Hazards 25(2):177–191, 2002. Kluwer Academic Publishers. Printed in the Netherlands. <http://www.springerlink.com/content/n160g0121800n742/fulltext.pdf>
- Zamecka A, Buchanan G (2000) Disaster risk management. Queensland Department of Emergency Services, Brisbane, p 115

Books and Reviews

- ABC (2006) Bushfire summer. Australian Broadcasting Commission, Melbourne. <http://www.abc.net.au/nature/bushfire/>
- Australian Bureau of Meteorology (2008) Protecting yourself and your home. Bushfire weather. BoM, Melbourne. <http://www.bom.gov.au/weather-services/bushfire/about-bushfire-weather.shtml>
- Bushnell S, Cottrell A, Spillman M, Lowe D (2006) Thuringowa bushfire case study. Understanding communities, Project BCRC Program C Community self-sufficiency for fire safety. Bushfire CRC, JCU CDS, Townsville
- CFA (2008) Country fire authority. Melbourne. <http://www.cfa.vic.gov.au>
- ESA (2008) Community education. Emergency Services Agency, Australian Capital Territory. http://www.esa.act.gov.au/esawebsite/content_esa/community_education/community_education.html
- FEMA (2008) Prepare for a wildfire. Federal Emergency Management Agency. U.S. Department of Homeland Security, Washington, DC. <http://www.fema.gov/plan-prepare-mitigate>
- Geoscience Australia (2005) Sentinel. Commonwealth of Australia, Canberra. <http://sentinel.ga.gov.au/acres/sentinel/index.shtml>
- Granger KJ, Smith DI (1995) Storm tide impact and consequence modelling: some preliminary observations. *Math Comput Model* 21(9):15–21 <http://www.em.gov.au/Emergency-Warnings/Pages/default.aspx>
- Roth W (1897) *Ethnological studies among the North West Central Queensland aborigines*. Queensland government, Brisbane
- Rural Fire Service (2008) Fire safety information. New South Wales rural fire service. NSW Government, Sydney. http://www.rfs.nsw.gov.au/dsp_content.cfm?CAT_ID=515. NSW Rural Fire Service, with information in 27 languages

Complex Dynamics of Air Traffic Flow

Banavar Sridhar and Kapil Sheth
NASA Ames Research Center, Moffett Field,
CA, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Complex Network Analysis
US Air Traffic Network
Future Air Traffic Scenarios
Concluding Remarks
Bibliography

Glossary

Air traffic flow Air Traffic Flow represents the distribution of air traffic over a region of space. Air traffic is undergoing major changes both in developed and developing countries. The demand for air traffic depends on population growth and other economic factors. Air traffic in the United States is expected to grow to two or three times the baseline levels of traffic in the next few decades. An understanding of the characteristics of the baseline and future flows is essential to design of a good traffic flow management strategy.

Complex networks A network connects components of a system. The connections and the number of components vary with the function of the network. It is extremely difficult to analyze and visualize the behavior of networks when the number of components in the system becomes large. There has been a major advance in our understanding of the behavior of networks with large number of components. Several theories have been advanced about the evolution of large biological and engineering

networks by authors in diversified disciplines like physics, mathematics, biology, and computer science.

Scale-free networks Several large biological and engineering networks exhibit a scale-free property in the sense that the probabilistic distribution of their nodes as a function of connections decreases slower than an exponential. These networks are characterized by the fact that a small number of components have a disproportionate influence on the performance of the network. Scale-free networks are tolerant to random failure of components but are vulnerable to selective attack on components.

Traffic flow management Safety limits the number of aircraft arriving at an airport or in a region of the airspace. Airspace Capacity, the maximum number of aircraft in a region, depends on the technology to keep aircraft separated by a safe distance and weather conditions. Airspace capacity decreases in the presence of severe weather and aircraft may have to be rerouted or delayed on the ground to maintain safety. The imbalance between airspace capacity and traffic flow demand leads to delays. Traffic flow management tries to maintain efficiency of the flows while not exceeding capacity limits.

Definition of the Subject

Civil Aviation is a vital sector of the US economy. Manufacturing of civil aircraft contributed \$59.9 billion to the US trade balance (exports less imports) in 2014. US airports handled 947 million passengers during 2016 (Federal Aviation Administration 2016). The National Airspace System (NAS) refers to all the hardware, software and people involved in managing air traffic in the United States. More than 42,000 commercial flights operated in the US airspace alone on a typical day at the present time. Although demand for air transportation has decreased from earlier estimates, system traffic expressed in revenue passenger miles (RPMs) is

projected to increase by 1.9% a year between 2018 and 2038 (Federal Aviation Administration 2018). This increase in demand will put a further strain on the airports and airspace resulting in delays. Federal Aviation Administration (FAA) estimated the annual costs of delays (direct cost to airlines and passengers, lost demand, and indirect costs) in 2017 to be \$26.6 billion (Bureau of Transportation Statistics 2018). To address the changes required in the air transportation system, FAA and the aviation industry have identified a set of priority capabilities for implementation (Federal Aviation Administration 2017). Similar activities are being pursued in Europe and Asia (Eurocontrol 2007).

Introduction

Air traffic in the United States has continued to grow at a steady pace since 1980, except for a dip immediately after the tragic events of September 11, 2001. There are different growth scenarios associated both with the magnitude and the composition of the future air traffic. The Terminal Area Forecast (TAF), prepared every year by the FAA, projects the growth of traffic in the United States (Federal Aviation Administration 2018). Both Boeing (2007) and Airbus (2002) publish market outlooks for air travel annually. Although predicting the future growth of traffic is difficult, there are two significant trends: (1) heavily congested major airports continue to see an increase in traffic and (2) the emergence of regional jets and other smaller aircraft with fewer passengers operating directly between non-major airports. The interaction between air traffic demand and the ability of the system to provide the necessary airport and airspace resources can be modeled as a network. The size of the resulting network varies depending on the choice of its nodes. It would be useful to understand the properties of this network to guide future design and development. Many questions, such as the growth of delay with increasing traffic demand and impact of the en route weather on future air traffic, require a systematic understanding of the properties of the air traffic network.

There has been a major advance in the understanding of the behavior of networks with a large number of components. Several theories have been advanced about the evolution of large biological and engineering networks by authors in diversified disciplines like physics, mathematics, biology, and computer science (Strogatz 2001). Several networks exhibit a scale-free property in the sense that the probabilistic distribution of their nodes as a function of connections decreases slower than an exponential. These networks are characterized by the fact that a small number of components have a disproportionate influence on the performance of the network. Scale-free networks are tolerant to random failure of components but are vulnerable to selective attack on components (Albert et al. 2000; Newman 2003).

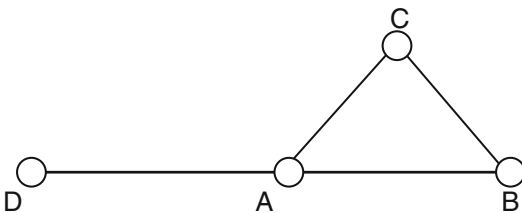
This chapter examines two network representations for the baseline air traffic system. A network defined with the 40 major airports as nodes and with standard flight routes as links has a characteristic scale: all nodes have 60 or more links and no node has more than 460 links. Another network is defined with baseline aircraft routing structure that exhibits an exponentially truncated scale-free behavior. Its degree ranges from 2 connections to 2900 connections, and 225 nodes have more than 250 connections. Furthermore, those high-degree nodes are homogeneously distributed in the airspace. A consequence of this scale-free behavior is that the random loss of a single node has little impact, but the loss of multiple high-degree nodes (such as occurs during major storms in busy airspace) can adversely impact the system. Two future scenarios of air traffic growth are used to predict the growth of air traffic in the United States. It is shown that a three-times growth in the overall traffic may result in a ten-times impact on the density of traffic in certain parts of the United States.

The chapter is organized as follows. The section on “[Complex Network Analysis](#)” provides a brief overview and terminology of the complex networks useful in the analysis of air traffic. This is followed by an application of the complex network methodology to the baseline air traffic system. The section on “[Future Air Traffic Scenarios](#)” considers different scenarios for the

evolution of air traffic in the United States during the next 25 years and looks for changes in the behavior of the network. An analysis of the future air traffic scenarios shows that a three-times growth in overall traffic can result in a tenfold impact on the density of highly connected nodes in certain parts of the United States. Thus, in addition to bottlenecks at major airports, the risk of airspace congestion calls for route restructuring and the introduction of new procedures and automation to increase airspace capacity. Concluding remarks and possible research in this area are discussed in the last section.

Complex Network Analysis

Complex systems have many agents or components interacting with each other, and their collective behavior is not a simple combination of the individual behavior. The pattern of interaction between agents can be studied as a network of connections between agents. Networks of many types are pervasive in modern society, and scientists from many fields are trying to broaden their understanding of the structure of the networks. A network is made up of basic components called either vertices or nodes. Each node is connected to other nodes in the system. The line connecting two nodes is referred to as a link or an edge. An edge is directed if it runs only in one direction and undirected if it runs in both directions. Degree is the number of edges connected to a node. Figure 1 shows a simple network with nodes A, B, C, and D and edges connecting them. The degrees of nodes A, B, C, and D are 3, 2, 2, and 1, respectively.



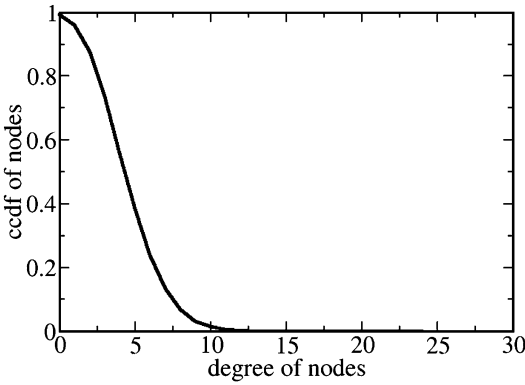
Complex Dynamics of Air Traffic Flow, Fig. 1 Nodes and edges in a network

The structure of the network strongly influences the functions performed by the network. It is possible to analyze networks of small sizes (less than 100) by drawing the picture of the network and analyzing its properties. The number of nodes in a complex engineering system, such as the worldwide web (Broder et al. 2000), can easily be as large as a few million nodes. The behavior of Random Graphs, graphs where nodes are connected to other nodes randomly, as the number of nodes becomes large has been studied extensively by mathematicians (Erdos and Renyi 1960). The behavior of Regular Networks, idealized systems where each component is identical, has been studied by several authors (Newman 2003). Networks describing real systems are neither random nor regular. The ability to model real systems and capture the behavior of key variables is extremely difficult. Recent developments in complex networks examine the statistical properties of large networks and help answer questions about the dynamics and stability of such networks. However, studies on the effects of structure on system behavior are still in their early stages.

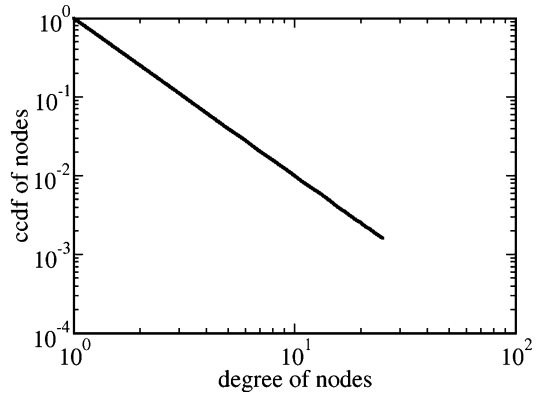
A major area of interest is the role played by the distribution of the degree of nodes in a network. Let p_k be the fraction of nodes in the network that has degree k . The degree distribution for the network can be computed by generating a histogram of p_k . The degree distribution for large random graphs, where each edge is present or absent with equal probability, can be modeled as a Poisson distribution. The distribution tends to peak for a small value of k and decays exponentially for large values of k . When using real data, to reduce the effect of noise in small datasets, the distribution of degree is expressed in terms of the complementary cumulative distribution function (ccdf), the probability of the number of nodes in the graph with degree greater than k . P_k is computed using the expression.

$$P_k = \sum_{i=k}^{\infty} p_i.$$

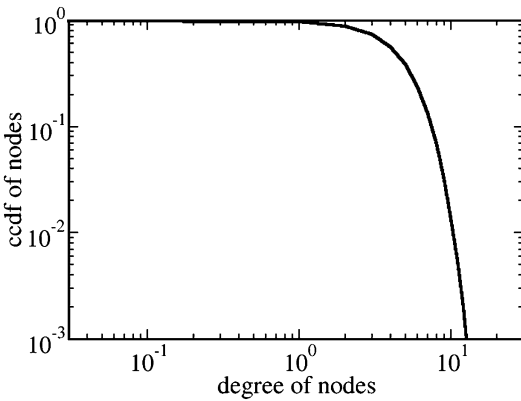
Figure 2 shows the ccdf of the Poisson distribution, $e^{-\lambda} \lambda^k / k!$, for $\lambda = 4$. Figure 3 shows the ccdf for the Poisson distribution on a logarithmic scale.



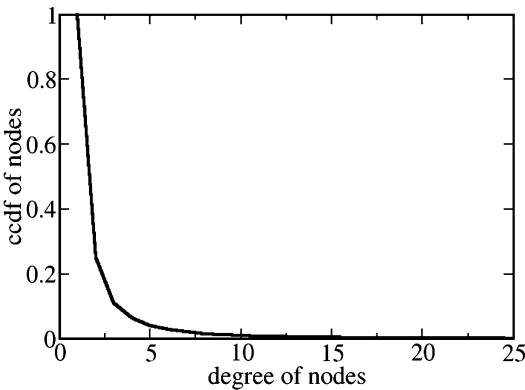
Complex Dynamics of Air Traffic Flow, Fig. 2 Poisson distribution



Complex Dynamics of Air Traffic Flow, Fig. 5 Power law distribution in logarithmic scale



Complex Dynamics of Air Traffic Flow, Fig. 3 Poisson distribution in logarithmic scale



Complex Dynamics of Air Traffic Flow, Fig. 4 Power law distribution

It has been observed that the degree distributions for some real-world networks, such as the Internet and biological networks, are highly skewed with tails several times longer than the mean. These real networks have a small number of nodes, or hubs, with a high level of connectivity. Hubs play an important role in influencing the properties of the network. Real networks exhibiting a small number of hubs are referred to as “scale-free” networks (Barabasi and Bonabeau 2003; Barabasi 2017). The distributions of degree for a number of networks, e.g., Internet, World Wide Web, collaboration network of mathematicians, show a power law in their tails: for some constant value of k . Figure 4 shows the ccdf for $\alpha = 2$. Power law distributions (Fig. 4) appear linear in a logarithmic scale (Fig. 5).

The appearance of hubs in scale-free networks is explained in terms of two behaviors observed both in people and real systems—growth and preferential attachment. There is a tendency for new growth to gravitate towards desirable locations. Generally, all the desirable locations already have some existing nodes. The preferential attachment mechanism leads to the creation of more powerful nodes or hubs. The various hubs compete to attract new nodes and absorb them, resulting in an increase of the number of links in the initial hubs.

Scale-free networks with power law distributions exhibit two important properties: (a) remarkable resistance to random failure of nodes and (b) extreme vulnerability to targeted attacks. The

functionality and performance of the network depends on the existence of the edges between pairs of nodes. The distance between the remaining vertices gets longer as nodes are removed from a network, and the removal of a certain number of nodes may result in a collapse of the entire system. The tolerance of networks to failures or removal of nodes varies with their level of resilience. The nodes from a network could be removed randomly or in a coordinated approach. Several studies have shown that scale-free networks are generally robust to random removal of nodes. However, they are less tolerant to the selective removal of hubs (Strogatz 2001).

The next section will examine the air traffic in the United States to see if it exhibits the properties of a scale-free network and draw analogies from the attack vulnerabilities of other complex networks. The methodology can also be used to understand the properties of the Next Generation Air Transportation System as it evolves over the next few decades (Pearce 2006).

US Air Traffic Network

The National Airspace System (NAS) refers to the collection of hardware, software, and people, including runways, radars, networks, FAA systems, airlines, etc., involved in Air Traffic Management (ATM) in the USA. It has been pointed out that the NAS should be modeled as a complex adaptive system (Donohue 2003). The NAS can be looked at as a network at several different levels (DeLaurentis et al. 2006). Some examples of networks involving components of the NAS are the route network of an airline connecting different airports, a network of sectors (geographical region used to monitor safe separation between aircraft) interconnected by geographical proximity, network of airline crew located in different cities and a surface network of ramps, runways and departure gates at an airport. The network components and links vary with the planning interval and the modeling problem. The analysis of baseline traffic uses actual data for a day from July 2006 and constructs two

Complex Dynamics of Air Traffic Flow, Table 1 Different types of air traffic management networks

Network	Nodes	Number of nodes	Edges
Airport network (AN)	40 major airports	40	Routes between 40 major airports
Airport and airspace network (AAN)	All airports and fixes along routes connecting the airports	8170	Routes between all airports

different types of networks for a typical day of operations in the NAS. The traffic data are processed using the Future Air traffic management Concept Evaluation Tool (FACET) simulation software (Sridhar et al. 2005) for conducting the network flow analysis.

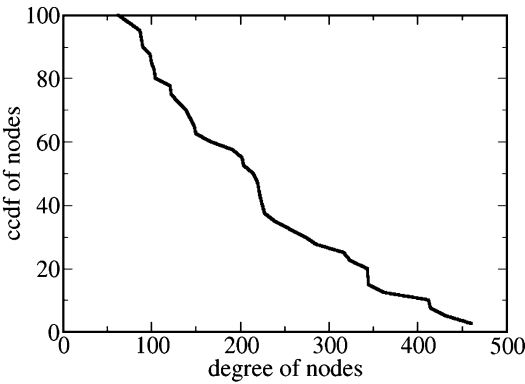
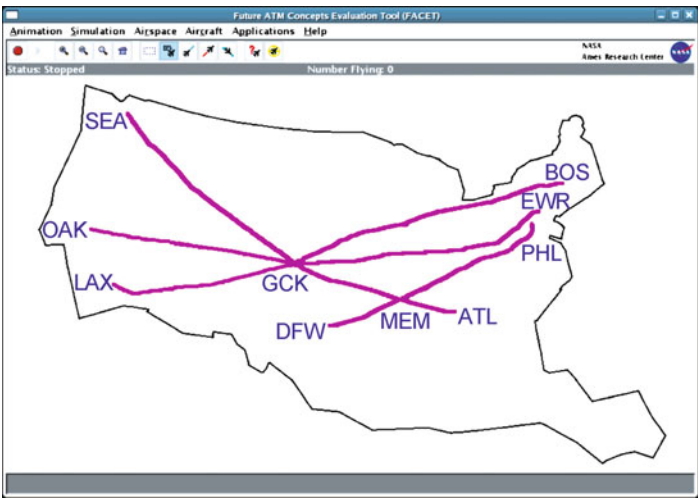
The major distinctions regarding the two air traffic networks are described in Table 1. The nodes in the Airport Network (AN) correspond to the 40 major airports in the USA. Today, a set of predefined alternative routes are used for flying between particular city pairs. Each node in network AN is connected to another node through one or more routes. The Airport and Airspace Network (AAN) includes all airports in the USA and routes connecting these airports. A route between two airports is defined by a series of geographical positions or fixes in the airspace. Each fix together with the airports represents a node in the AAN network.

Computation of the Degree of a Node and the Distribution Function

The computation of the degree of a node and the distribution function is illustrated by an example. The FACET simulation software generates the departure airport, arrival airport, and the fixes through which the aircraft travels en route based on air traffic data. Figure 6 shows the intersection of different routes connecting airports through common fixes. In the figure, aircraft traveling from Seattle (SEA) to Atlanta (ATL), aircraft traveling from Oakland (OAK) to Newark (EWR),

Complex Dynamics of Air Traffic Flow,

Fig. 6 Example of intersection of routes at common fixes



Complex Dynamics of Air Traffic Flow,
Fig. 7 Cumulative distribution function for nodes (40 major airports) in network AN

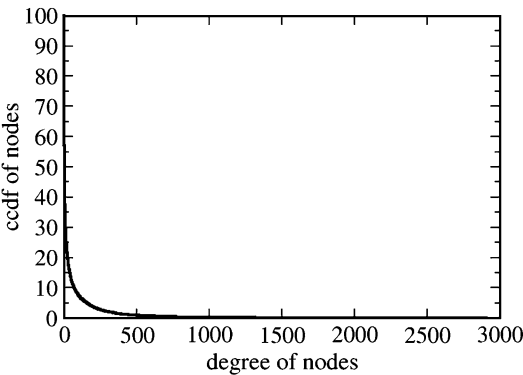
and aircraft traveling from Los Angeles (LAX) to Boston (BOS) pass through the fix Garden City, Kansas (GCK). The degree of GCK is equal to the total number of aircraft traveling along these three routes during the day. The degree of node GCK is three if a single aircraft travels between each of these three pairs of airports. Similarly, aircraft traveling from Dallas (DFW) to Philadelphia (PHL) and Seattle to Atlanta pass through the fix Memphis, TN (MEM). The degree of fix MEM is two if a single aircraft travels between each of these city pairs. The distribution and the ccdf can be computed easily given the degree associated with all the airports and en route fixes and are described in the next section.

Behavior of the Degree of Nodes in Air Traffic Management Network

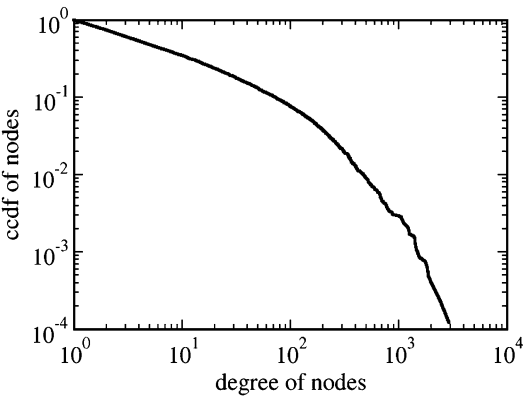
The ccdf for the nodes in network AN is shown in Fig. 7. The figure shows the probability that an airport in the network has more than a certain number of routes originating or terminating at the airport. The number of nodes at all airports exceeds 60 and Chicago O'Hare airport has the maximum number (460) of connections. The distribution of nodes in the airport network does not show scale-free behavior. The airport connection distribution in the USA is similar to the distribution of the connections between the networks of world airports (Amaral et al. 2000). A recent study analyzing global air traffic lists London Heathrow, Chicago O'Hare, Frankfurt, Amsterdam, Toronto, Los Angeles, Atlanta, Singapore, Paris Charles De Gaulle, and Jakarta as the top ten airports with most connections (Anonymous 2018).

Figure 8 shows the ccdf of the nodes in network AAN. The ccdf of the nodes in network AAN shows a rapid decay of the percentage of nodes with the degree of the node. However, it has 225 nodes with more than 250 links, indicative of high volumes of traffic through these nodes. Figure 9 shows the logarithmic behavior of the ccdf of the nodes in network AAN.

The nodes of network AAN initially follow the power law curve similar to several scale-free networks (Newman 2003). However, towards the end of the tail, the distribution shows a deviation



Complex Dynamics of Air Traffic Flow,
Fig. 8 Distribution of nodes in AAN with baseline traffic



Complex Dynamics of Air Traffic Flow,
Fig. 9 Logarithmic behavior of nodes in network AAN

from the power law behavior indicating a limit on the number of nodes that have a high number of connections. The limit on the growth of the large hub nodes, in the distribution of the ATM network, is due to the constraints imposed on traffic demand by the location of the cities, economic development, and government policies.

The existence of hub and secondary airports characterizes baseline air traffic operations (Bonnefoy and Hansman 2004). The network analysis provides an additional ability to study traffic behavior in the en route airspace. Figure 10 shows the geographical distribution of the nodes with degree higher than 250 in network AAN. These nodes will be referred to as 250G nodes. 250G nodes represent 2.7% of the

total nodes in the network. Table 2 shows the number of 250G nodes in different traffic control regions, referred to as Centers. The Centers, Chicago (ZAU), Boston (ZBW), Atlanta (ZTL), Kansas City (ZKC), Washington, DC (ZDC), Indianapolis (ZID), Jacksonville (ZJX), Los Angeles (ZLA), Cleveland (ZOB), and New York (ZNY), each have more than ten of the 250G nodes. The Centers in the western part of the United States have fewer 250G nodes indicative of the lower traffic density in these Centers. It is informative to express the distribution of the 250G nodes in units of number of nodes per 10,000 square nautical miles (10Ksqnm). Using this measure of nodal traffic density, ZNY has the highest nodal density with six 250G nodes per 10Ksqnm. Table 3 shows the nodal traffic density for the 20 Centers in the continental USA.

The behavior of the degree of the nodes in an ATM network should not be surprising, since the ATM network has evolved to serve population densities in the USA. Network analysis helps to visualize and quantify the characteristics of the network.

Resilience of Air Traffic Management Networks

The tolerance of complex networks to random and targeted failures depends on their network structure. As observed earlier, a scale-free network is tolerant to random failure, since the hubs are few and the chance of a hub being selected randomly is low. However, the same network may be prone to targeted attacks on a small percentage of vital nodes. In air traffic management networks, weather can be regarded as an agent of attacks on the system.

Convective weather is a major source of uncertainty in ATM networks. One effect of severe weather is to make airspace unavailable for the flow of air traffic. The removal of airspace may result in the failure of hubs and increase the average path length in the network. The impact of weather on ATM system performance appears as delay (Sridhar and Swei 2006). The tolerance of the ATM network to weather depends on the geographical distribution of the weather and the coverage of nodes with high degree.



Complex Dynamics of Air Traffic Flow, Fig. 10 Geographical distribution of 250G nodes in network AAN with baseline traffic and severe weather polygons

Future Air Traffic Scenarios

National and international projections of traffic growth in 2006 predicted a tripling of passengers by 2025 (Boeing- Current Market Outlook 2007). There may be increased traffic due to the growing presence of on-demand air taxis and unmanned air vehicles. It is estimated that 5000 microjets may be operational by 2010 and 13,500 by 2022 (Teal Group 2003). The TAF provides forecasts for airports in the NAS. The forecast for the major airports in the United States receives more detailed modeling that takes into account local economic conditions and airline costs. TAF is the basis for most aviation demand forecasts. The TAF data published in March 2006 provide traffic growth rates for the period 2005–2021. The AvDemand model (Huang et al. 2006) provides three times (3X) baseline traffic scenarios starting with TAF and assuming slightly different conditions for traffic growth beyond 2025. The Transportation System Analysis Model (TSAM) (Viken et al. 2006) predicts future demand travel based on

demographics and population at the county level. It provides a complimentary alternative to the FAA forecast. The results in this study are based on data generated by AvDemand.

Two future scenarios are considered. These two scenarios and the baseline traffic provide a baseline to compare the impact of future traffic changes on the ATM system. The initial values of aircraft routes and schedules used in the extrapolation are based on traffic data for a day in May 2002. Assuming traffic growth from TAF and the nominal traffic schedule, the Fratar algorithm (Viken et al. 2006) can be used to create future daily total number of flights between each origin and destination pair of the baseline schedule. The traffic growth scenarios using TAF airport growth out to 2025 predict a future NAS-wide traffic growth by a factor of 1.49. To achieve three-times baseline day traffic (3X) scenarios assumed in future planning (Borener et al. 2006), consider two different scenarios for post-2025 traffic growth. The compound extrapolation approach grows traffic until it reaches 3X by assuming the

Complex Dynamics of Air Traffic Flow, Table 2 Distribution of 250G nodes by Centers under baseline, 2017 and future traffic scenarios

Center Name	Symbol	Baseline	2017	3XH	3XC
Albuquerque	ZAB	7	7	70	78
Chicago	ZAU	15	6	67	70
Boston	ZBW	16	7	77	87
Washington, D.C.	ZDC	22	40	134	143
Denver	ZDV	8	3	51	68
Dallas Fort Worth	ZFW	3	2	59	68
Houston	ZHU	8	9	60	72
Indianapolis	ZID	15	17	71	71
Jacksonville	ZJX	14	24	71	81
Kansas City	ZKC	11	11	58	50
Los Angeles	ZLA	16	10	102	119
Salt Lake City	ZLC	5	2	29	40
Miami	ZMA	4	8	44	54
Memphis	ZME	10	8	38	42
Minneapolis	ZMP	6	4	46	53
New York	ZNY	14	18	80	88
Oakland	ZOA	10	5	57	67
Cleveland	ZOB	13	8	90	89
Seattle	ZSE	5	4	35	41
Atlanta	ZTL	15	9	59	62
	Total	217	202	1298	1443

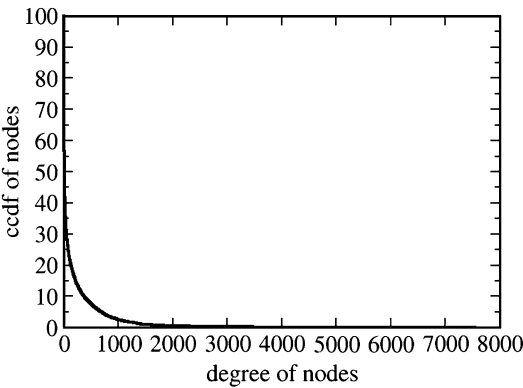
TAF airport growth rates for traffic beyond 2025. As an alternative, the homogeneous extrapolation approach assumes the same growth rates at all airports until the traffic level reaches 3X. These two scenarios will be referred to as 3X Compound (3XC) and 3X Homogeneous (3XH) scenarios in the rest of the chapter. The scenarios are processed to compute the changes in the network properties of future ATM systems.

Behavior of Future ATM Networks

The two scenarios described earlier can be used to study the characteristics of the 3X traffic under the two assumptions. The properties of future ATM networks can be derived similar to the properties of the baseline ATM networks. The computations can be used to compare the geographical distribution of the hubs compared to the distribution today. Figures 11 and 12 show the distribution of the nodes in the AAN network for the 3XC traffic

Complex Dynamics of Air Traffic Flow, Table 3 Nodal density of 250G nodes by Centers under baseline, 2017 and future traffic scenarios

Center	Area	Baseline	2017	ρ_{3XH}	ρ_{3XC}
ZAB	18.04	0.39	0.39	3.88	4.32
ZAU	7.60	1.97	0.79	8.82	9.21
ZBW	11.62	1.38	0.6	6.63	7.49
ZDC	12.69	1.73	3.15	10.56	11.27
ZDV	20.18	0.40	0.15	2.53	3.37
ZFW	12.32	0.24	0.16	4.79	5.52
ZHU	27.64	0.29	0.33	2.17	2.61
ZID	7.07	2.12	2.4	10.04	10.04
ZJX	14.55	0.96	1.65	4.88	5.57
ZKC	13.37	0.82	0.82	4.34	3.74
ZLA	13.55	1.18	0.74	7.53	8.78
ZLC	31.57	0.16	0.06	0.92	1.27
ZMA	28.66	0.14	0.28	1.53	1.88
ZME	10.75	0.93	0.74	3.53	3.91
ZMP	28.82	0.21	0.14	1.60	1.84
ZNY	2.43	5.76	7.41	32.92	36.21
ZOA	13.54	0.74	0.37	4.21	4.95
ZOB	6.74	1.93	1.19	13.36	13.21
ZSE	19.64	0.25	0.2	1.78	2.09
ZTL	9.18	1.63	0.98	6.43	6.75
Total	309.96	0.70	0.65	4.19	4.66

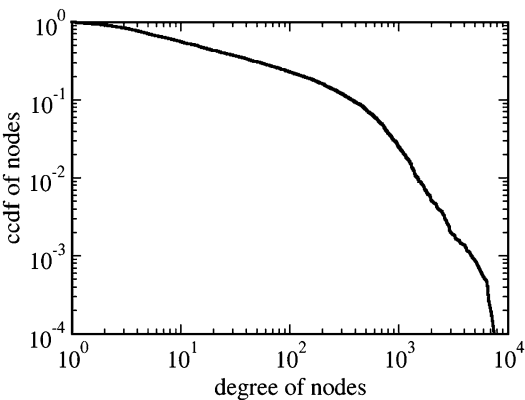


Complex Dynamics of Air Traffic Flow, Fig. 11 Distribution of nodes in NAA using 3X Compound Scenario

scenario. The results are similar for the 3XH traffic scenario.

The future ATM networks, under both scenarios, show exponentially truncated scale-free behavior similar to the baseline ATM network. The distribution of the hubs will have an impact

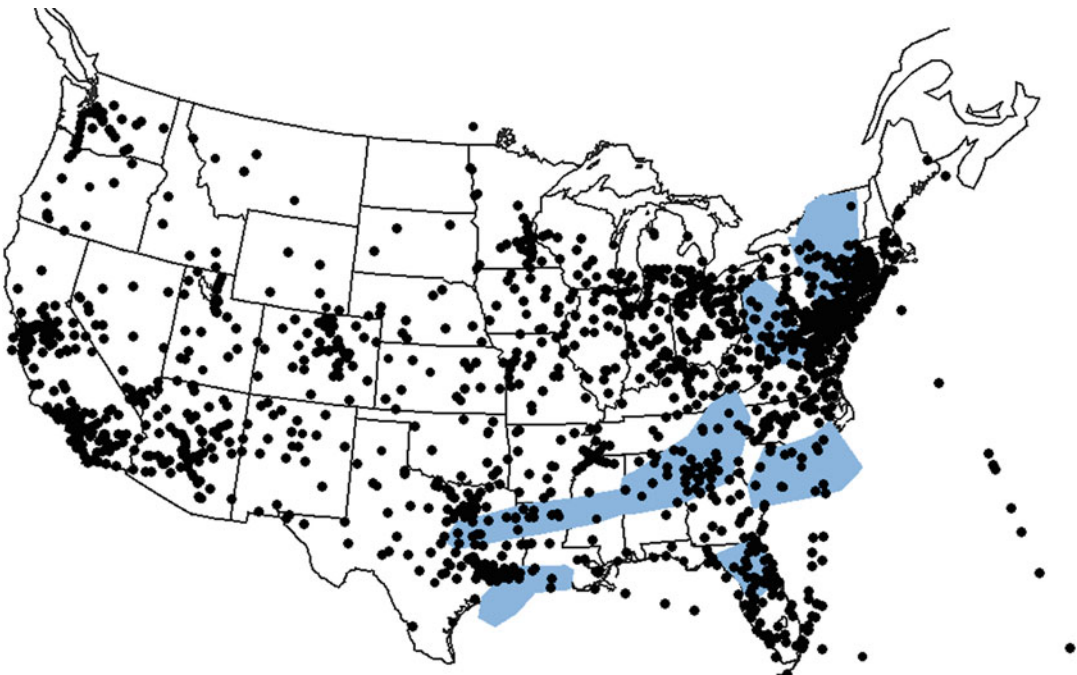
on the growth of delay in future ATM networks subjected to severe weather. The vulnerability of the network to reduction in capacity caused by certain weather patterns will be significant and disruptive to the operation of the system. The impact of the weather on future ATM networks will be described subsequently in the paper.



Complex Dynamics of Air Traffic Flow, Fig. 12 Logarithmic behavior of nodes in network NAA under 3X Compound traffic

The geographical distribution of the 250G nodes under 3XC is shown in Fig. 13. Table 2 shows the same distribution by Centers for baseline traffic and 3X scenarios. An examination of Table 2 is helpful in drawing distinctions between baseline day traffic and 3X demand traffic. The 3X traffic demand creates close to a six-fold increase in the 250G traffic nodes. Earlier, it was noted that ZNY has the highest 250G traffic nodal density. Under the 3X demand, as can be seen from Table 3, ZNY nodal density of 2006, 5.76, is equaled or exceeded by a majority of the 20 Centers. Even more alarmingly, the nodal density of 250G nodes in ZNY is six times the baseline value, and Cleveland and Washington Centers have twice the nodal density of ZNY in 2006. The 3X traffic also gives rise to nodes with even more connections. The hubs, 500G, 750G, and 1000G, are defined similarly to the 250G hub. The number of 250G, 500G, 750G, and 1000G hubs under different traffic conditions are shown in Table 4.

The actual growth of traffic during the last 10 years has been closer to 2% annually (Bureau



Complex Dynamics of Air Traffic Flow, Fig. 13 Geographical distribution of 250G nodes in network N5 with 3X Compound traffic and severe weather polygons

Complex Dynamics of Air Traffic Flow, Table 4 Total number of nodes for different traffic scenarios

Types of nodes	Current	2017	3XH	3XC
250G	225	212	1312	1468
500G	72	57	620	806
750G	33	29	304	452
1000G	22	21	165	262

of Transportation Statistics 2018), smaller than the 3X growth used in the analysis. Tables 2, 3, and 4 provide the distribution of the 250G nodes in different Centers, the nodal density in different centers, and the number of 250G, 500G, 750G, and 1000G nodes using traffic data for the Year 2017. A comparison of the results between baseline year and 2017 shows that the traffic density is similar and critical regions in the airspace continue to experience congestion and delays.

Impact of Weather on ATM Networks

Another way to view the growth of traffic is to compare how similar weather patterns may affect baseline traffic and 3X traffic. The geographical distribution of the 250G nodes shown in Fig. 13 is based on traffic during July 2, 2006. This was a calm weather day with a total NAS aggregate delay of 11,997 min (Federal Aviation Administration 2004). The traffic on July 2, 2006, will be assumed as traffic unaffected by weather in this discussion (Sridhar and Swei 2006).

Whenever there is severe weather in the NAS, airspace capacity is reduced and traffic is rerouted or held on the ground causing delays in the system. The Collaborative Convective Forecast Product (CCFP) is a model of severe weather activity and the areas marked blue in Fig. 10 show the CCFP for July 13, 2006. On July 13, 2006, the NAS experienced a significant total delay of 219,350 min with ZNY, ZDC, and ZTL Centers contributing a delay of 116,654, 49,400, and 23,822 min, respectively. Next, as shown in Fig. 13, the same severe weather is overlaid in blue polygons on the geographical distribution of the 250G nodes under 3XC demand. Table 5 shows hub nodes affected by the weather today and in the future.

Complex Dynamics of Air Traffic Flow, Table 5 Total number of nodes affected by severe weather

Types of Nodes	Baseline	3XH	3XC
250G	34	191	207
500G	9	96	118
750G	4	49	67
1000G	3	23	41

A greater number of hubs, between six and ten times, are affected by the same weather pattern under future traffic scenarios than today. If one considers the nonlinear growth of delay, in regions such as ZNY, ZDC, and ZTL where the demand on the airspace is close to capacity, the increased density of high traffic nodes in these regions will result in much larger delays compared to 2006.

Concluding Remarks

Air traffic in the United States can be modeled as a network to understand the impact of the predicted growth in the demand on the performance of the system. It is demonstrated that the air traffic network with baseline en route flight plan intersections as nodes and with the flight plans as links shows scale-free properties typical of several large engineering and biological networks. A consequence of this property is the nonlinear growth of traffic in certain regions of the United States. A preliminary analysis indicates that a three-times growth in the overall traffic may result in a ten-times impact on the density of traffic in certain parts of the United States. The air traffic system currently experiences significant delay during periods of severe weather activity. The impact of weather of the same severity will be magnified several times, especially in the northeastern parts of the United States, leading to lower system performance. Recent research analyzes the vulnerability of ATM networks to computer failures and targeted attacks (Roy et al. 2017). The actual growth of traffic during the last 10 years has been closer to 2% annually (Bureau of Transportation Statistics 2018), smaller than the 3X growth used in the

analysis. However, the smaller growth rate provides Airlines, FAA, and other organizations more time to execute policies to maintain the safety and improve the efficiency of air traffic. Research must be conducted to determine whether this risk to system performance can be mitigated through restructuring routes or by introducing new operational concepts, such as automation assistance to controllers to increase airspace capacity (Schroeder 2009; Sridhar and Chatterji 2015). The network analysis described in the chapter can be used to guide the development of various traffic flow management concepts to increase the efficiency of air traffic systems.

Bibliography

- Airbus (2002) Global market forecast 2001–2002. Reference CB 390.0008/02. Airbus S.A.S, Blagnac
- Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382
- Amaral LAN, Scala A, Barthélemy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci U S A* 97(21):11149–11152
- Anonymous (2018) OAG Megahubs international index. The world's most internationally connected airports
- Barabasi AL (2017) *Network science*. Cambridge University Press, Cambridge
- Barabasi AL, Bonabeau E (2003) Scale-free networks. *Sci Am* 288:50–59
- Boeing- Current Market Outlook (2007) <http://www.boeing.com/commercial/cmo>
- Bonnefoy PA, Hansman RJ (2004) Emergence and impact of secondary airports in the United States. In: 4th AIAA aviation technology, integration and operations conference, Chicago
- Borener S, Carr G, Ballard D, Hasan S (2006) Can NGATS meet the demands of the future? *J Air Traffic Control* 48:34–38
- Broder A et al (2000) Graph structure in the web. *Comput Netw* 33:309–320
- Bureau of Transportation Statistics (2018) 2017 Annual and December U.S. airline traffic data. <https://www.bts.dot.gov/newsroom/2017-annual-and-december-us-airline-traffic-dataBureau>
- DeLaurentis D, Han E, Kotegawa T (2006) Establishment of a network-based simulation of future air transportation concepts. In: 6th AIAA aviation technology, integration and operations conference, Wichita
- Donohue G (2003) Air transportation is a complex adaptive system: not an aircraft design. In: AIAA/ICAS international air and space symposium and exposition Dayton
- Erdos P, Renyi A (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Sci* 5:17–61
- Eurocontrol (2007) An assessment of air traffic management in Europe during the calendar year 2006: performance review report. Eurocontrol, Brussels
- Federal Aviation Administration (2004) Order 7210.55C: operational data reporting requirements. U. S. Department of Transportation
- Federal Aviation Administration (2016) The economic impact of civil aviation on the U.S. Economy. https://www.faa.gov/air_traffic/publications/media/2016-economic-impact-report_FINAL.pdf
- Federal Aviation Administration (2017) NextGen priorities: joint implementation plan update including the Northeast Corridor. <https://www.faa.gov/nextgen/media/NGPriorities-2017.pdf>
- Federal Aviation Administration (2018) Terminal area forecast summary, fiscal years 2017–2045. http://www.faa.gov/data_research/aviation/taf
- Huang AS, Schleicher D, Hunter G (2006) Future flight demand generation Tool. In: 6th AIAA aviation technology, integration and operations conference, Wichita
- Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45:167–256
- Pearce RA (2006) The next generation air transportation system: transformation starts now. *J Air Traffic Control*:7–10
- Roy S, Xue M, Sridhar B (2017) Vulnerability metrics for the airspace system. Twelfth USA/Europe air traffic management research and development seminar (ATM 2017), Seattle
- Schroeder JA (2009) A perspective on NASA Ames Air Traffic Management Research. AIAA-2009–7054. In: AIAA aviation, technology, integration, and operations (ATIO) conference and aircraft noise and emissions reduction symposium (ANERS), Hilton Head
- Sridhar B, Chatterji G (2015) *Air traffic management*. In: John Wiley Encyclopedia of electrical engineering. Wiley
- Sridhar B, Swei SM (2006) Relationship between weather, traffic and delay based on empirical methods. In: 6th AIAA aviation technology, integration and operations conference, Wichita
- Sridhar B, Sheth K, Smith P, Leber W (2005) Migration of FACET from simulation environment to dispatcher decision support system. In: Proceedings of digital avionics systems conference, Washington
- Strogatz SH (2001) Exploring complex networks. *Nature* 410:268–276
- Teal Group (2003) World unmanned aerial vehicle systems, market profile and forecast, Fairfax, VA
- Viken J, et al (2006) NAS demand predictions, transportation systems analysis model (TSAM) compared with other forecasts. In: 6th AIAA aviation technology, integration and operations conference, Wichita

Index

A

Acceleration-based models, 649, 654
 HTM, 654–655
 limitations of, 656
 SFM, 655–656
Acceleration time delay three-phase traffic flow model (ATD model), 576
Action points, 571
Active inference framework, 571
Activity-based approach, 613, 628–629
Activity Mobility Simulator (AMOS), 634–635
Actuated control, 132
Adaptation, 626
Adaptive cruise control (ACC), 42, 45–48, 347, 349
 classical (standard), 349
 in framework of three-phase theory, 356
 three-phase theory, 54–56
Advanced data collection methods and devices, 633
Advanced driver assistance systems (ADAS), 187
Advanced sensor technologies, 186–187
Advanced traveler information systems (ATISs), 80, 116
Airport and Airspace Network (AAN), 745
Air traffic, 16, 742
 complex network analysis, 743
 flow, 1, 741
 future scenarios, 748–751
 US air traffic network, 745–747
Alam Penn State Emergency Management model, 636
ALBATROSS, 636, 637
ALINEA, 179
Alternative network routes, 22
ANCONA, 179
 on-ramp metering, 45
Arterial systems, progression models
 arterial signal timings, 141
 closed loops, 141
 closed network, 141
 mixed-integer linear programming (MILP), 142
 open network, 141
Artificial neural networks (ANNs), 535–536
Asymmetric simple exclusion process (ASEP), 659, 662
Attitudes, 711
AURORA, 637
Australia, 702

Authority constraints, 628
Authority's Perceived Need to Evacuate (APNE), 707
Automated driving vehicles, 42, 45–48, 245
 based on three phase theory, 53–63
 classical model of, 46–48
 deterioration of performance of traffic system, 49–53
 dynamic behavior of, 56–57
 probability of traffic breakdown, 45–49
 without fixed desired time headway to preceding vehicle, 54–56
Automated Highway System (AHS), 188
Automatic driving, 33, 42
Automatic traffic control
 efficiency, 172
 fuel consumption, 173
 network-oriented, 169–172
 network reliability, 173
 safety, 172–173
Autonomous driving, 1, 15, 33, 42, 343
 in framework of three-phase theory, 343
 framework of three-phase traffic theory, 1, 343
Auto-regressive integrated moving average (ARIMA), 540
Average arrival traffic rate, 58
Aw-Rascle macroscopic traffic flow model, 233

B

Backcasting, 617–618
Bando optimal velocity model, 232
Bandwidth problem, 131
 maximization, 142–144
 network problem, 147–149
 variable, 144–147
Behavior, 711, 714
 paradigm, 625
 travel-demand models, 627
Birth-and-death process, 299
Blue-Adler model, 660
Bottleneck, 1, 5, 7, 10, 21, 195, 343, 387, 650, 671
 capacity, 226
 critical strength of heavy, 470
 effective location of, 439
 effectual, 285, 387, 439
 flow, 682

- Bottleneck (*cont.*)
 heavy, 390, 467
 induced traffic breakdown at, 22, 196
 moving, 391, 439
 strength, 467
 threshold strength, pinch region existence, 470
- Braess network, 86
- Brake light models
 comfortable driving model, 329
 desire time gap brake light model, 333
 modified comfortable driving model, 331
- Breakdown minimization (BM) principle, 1, 8
 alternative network routes, 64
 application of, 64–65
 definition, 21, 28, 63
 motivation, 63
 network capacity, 65–67
 network throughput maximization approach, 65–66
 vs. Wardrop's equilibria, 68–70
 zero breakdown probability, 64, 65, 69
- Brilon's stochastic highway capacity, 210
- Buses
 complex dynamics of delay, 595, 598–600, 603, 606–609
 dynamic model of shuttle bus system, 595
 freely passing buses, 604–606
 increase (decrease) of, 608–610
 no passing buses, 606–608
 schedules, 594
 shuttle bus systems, 594
 single bus, 596, 601, 608
- C**
- Cairns City, North Queensland, 722
- Camera traffic detection, 174
- Capability constraints, 628
- Capacity, 671
- Car following behavior, 176
- Car-following model, 79, 100, 102, 315, 338
- Car-following problem
 driving simulator experiments, 576–581
 dynamical trap model, 568, 581–585
 following-the-leader models, 573
 Newtonian approach, 572–573
 social force model, 572
 and three-phase traffic theory, 574–576
- Car sharing programs, 622
- Catch effect, 212, 444, 460
- Cellular automata (CA) models, 176, 313, 573, 649, 659–660
 advantages, 315
 Blue-Adler model, 660
 comfortable driving model, 329–331
 desire time gap brake light model, 333–336
 floor fields, 660–662
 future research, 338
 Gao model, 324–327
 Kerner-Klenov-Schreckenberg model, 321–324
 Kerner-Klenov-Wolf model, 320–321
 lattice-gas models, 662
 limited deceleration model, 335–338
 modified comfortable driving (MCD) model, 331–333
 NaSch model, 317–318
 RCA, 662
 slow-to-start model, 318–320
 traffic flow model, 1, 14
 two-state model, 327–329
 velocity effect model, 317
- Celosity-dependent-randomization (VDR) rule, 318
- Classical traffic and transportation theories, failure of, 27–28
- Classic General Motors (GM) model
 achievements of, 238
 Aw-Rascle macroscopic model, 233
 Bando optimal velocity model, 232
 free flow instability, 235–236
 intelligent driver model, 232
 Krauß's stochastic microscopic model, 234–235
 Nagel-Schreckenberg (NaSch) stochastic cellular automaton (CA) model, 233–234
 Newell optimal velocity (OV) model, 232
 Payne-model, 232–233
 Wiedemann's psychophysical model, 233
- Classic Lighthill-Whitham-Richards (LWR) model
 achievements of, 228–229
 Daganzo's cell-transmission model, 227
 disadvantages, 229–231
 traffic congestion, 226–227
- Classification and regression tree (CART) algorithms, 115
- Cognition, 711
- Colored caps, 690
- Combined map model, 593, 603–604
- Comfortable driving model, 329
- Communication safety triangle (CST), 703, 704
- Community, 1, 701, 706
- Community empowerment, 701
- Community preparations and evacuation plan (CPEP), 727
- Community resilience, 707
- Community safety groups, 701, 725–728
- Complex dynamics
 of bus, tram and elevator delays, 595, 598–600, 603, 606–609
- Complexity and systems science (CSS), 702
- Complex network, 741
 analysis, 743–745
- Comprehensive Econometric Microsimulator for Daily Activity-travel Patterns (CEMDAP) model, 637
- Comprehensive Modal Emissions Model (CMEM), 114
- Computational process models (CPM), 633, 635
- Conceptual models, 624
- Congestion (traffic), 4, 13, 15, 21, 26, 33, 80, 195, 203, 230, 387
 control approach, 1–2, 13, 21–22, 45, 179
 diagram of, 389

- expanded pattern, 389
 - general pattern, 389
 - at heavy bottleneck, 390
 - strong, 458
 - Controller Area Network (CAN), 521
 - Convective weather, 747
 - Cooperative social systems, 561
 - Cooperative Vehicle-Infrastructure Systems (CVIS), 187–188
 - Cost function, 182
 - Coupling constraints, 628
 - Crowd, 2, 649, 671
 - disaster, 671
 - dynamics, 16
 - Cyclic flow profiles, delay-based models
 - GO-pattern/saturation flow pattern, 149
 - IN-Pattern/arrival flow pattern, 149
 - optimization model, 150–151
 - OUT-Pattern/departure flow pattern, 149–150
 - TRANSYT, 149
 - Cyclone Ingrid, 712
 - Cyclone Larry, 709, 725
 - Cyclone Tracy, 730
- D**
- Daganzo's cell transmission model, 227–228
 - Daily Activity and Travel Scheduling (DATS), 634
 - Data mining models, 15, 501, 533
 - Decision-based models, 649
 - See also* Rule-based models
 - Dedicated lanes, 183
 - Delay-based models, cyclic flow profiles, 131
 - GO-pattern/saturation flow pattern, 149
 - IN-Pattern/arrival flow pattern, 149
 - optimization model, 150–151
 - OUT-Pattern/departure flow pattern, 149–150
 - TRANSYT, 149
 - Delayed map model, 593, 598–600
 - Demand-responsive models
 - OPAC, 153–159
 - SCOOT, 151–153
 - strategies, 151
 - Density waves, 677
 - Desire time gap brake light (DTGBL) model, 333
 - Detector, 131
 - traffic flows from, 131, 152
 - vehicle, 132
 - Deterministic chaos, 593
 - Deterministic traffic breakdown, 293
 - Disaster, 2
 - definition, 701, 706
 - lead time, 701
 - management, 702 (*see also* Evacuation processes)
 - threat, 701
 - Discrete choice models, 627
 - Dissolving GP (DGP), 458
 - Downstream jam front, 391
 - 2D region of states of synchronized flow, 344
 - Driver behavior, 15
 - Driver experiment, 432
 - on circular road, 431
 - Driver routing behavior, 80
 - Driver time delay in acceleration, 297
 - Driver travel decision behavior modeling
 - intelligent transportation systems (ITS) program, 83
 - theory, 84
 - traffic assignment, 83
 - Dynamical trap, 559, 568
 - car-following problem, 568–569
 - fuzzy headway-speed diagram, 568
 - model, 576
 - two-component approach, 569
 - Dynamic floor field, 660
 - Dynamic model of shuttle bus system, 595
 - Dynamic planning, 613
 - practice, 614–618
 - Dynamic/time-dependent origin-destination estimation
 - dynamic formulations, 109–110
 - Nguyen–Dupuis network, 113
 - O-D generation problem, 105–106
 - solution algorithms, 110–113
 - static formulations, 107–109
 - synthetic O-D and trip distribution formulations, 106–107
 - trip-based approach, 107
 - trip distribution process, 107
 - Dynamic traffic assignment (DTA) system, 2, 4, 13, 22, 79
 - approach, 81
 - assignment, and assessment, 81
 - CONTRAM, 97
 - definition, 80
 - framework, 81
 - INTEGRATION traffic simulation, 98
 - macroscopic modeling approaches, 100–101
 - mesoscopic modeling approaches, 101
 - microscopic modeling approaches, 101–102
 - O-D matrix, 81
 - on-line application, 81
 - and routing framework, 82
 - solution approach, 96–99
 - synchronized flow phase, 100
 - three-phase traffic flow theory, 100
 - user equilibrium vs. system optimum, 87–88
 - wide moving jam, 100
 - Dynamic traffic control, 168
 - Dynamic traffic routing
 - assignment approach, 89
 - Bechmann's transformation, 91
 - definition, 80
 - equivalency conditions, 91
 - first-order conditions, 92
 - formulations, 89
 - Hessian matrix, 93
 - Lagrange multiplier, 95

Dynamic traffic routing (*cont.*)

- link-path incident matrix, 90
- link performance functions, 93
- mathematical formulations, 89–91
- uniqueness condition, 93
- user equilibrium, 91–93

Dynamic travel time estimation

- artificial neural network (ANN) techniques, 103
- automatic vehicle identification (AVI) data, 103
- genetic algorithm (GA) approach, 105
- moving average (MA) technique, 103
- probe vehicle technologies, 102
- reformulation-linearization technique, 105
- spot speed measurement systems, 102

E

Economic Commission for Asia and the Pacific (ECAP), 712

Effective risk communication, 701

Elevators, complex dynamics of delay, 595, 598–600, 603, 606–609

Elimination by aspects (EBA) model, 626

Emergency, 701

Empirical fundamental of transportation science, 2, 22, 29–33

Empirical induced traffic breakdown, 216
minimum highway capacity, 218

Empirical traffic breakdown, 22, 31

Engrams, 571

Enhance system performance, 116–118

Evacuation processes, 2, 16, 671, 701

communication, 706–707

Community Safety Group, 727–728

crowd disasters, 686–687

definition, 702–703

delivering real-time warnings, 728–729

description, 685

disaster, 706

disaster impacts, 722–727

effective risk communication, 709–713

guidelines and legal requirements, 685–686

hazards, 704–705, 719

images and warning maps, 733

institutional barriers to greater community self-help, 720–722

integrating theory and implementation, 713–719

issues, 704

linking people with message, 731–732

mathematical modeling, 707–709

media, 713, 718–719, 729

modeling risk communication, 733

new cultures to hazard zones, 733–734

playing the odds, 734

precautionary approach, 712

preparations, 703

risk communication, 712

risk communication theory and residents, 728

sirens or not, 713

spreading the warning, 734–735

Stern Report, 714

storytelling links, 719

timeliness of warnings, 730–731

traditional weather warning, 732

triggering action, 729–730

web delivery, 719

words and images, 732–733

Evacuation triggers (ET), 707

Event-driven model, 570

Exit routes, 707

Expanded congested pattern (EP), 389, 460

Extended circle map model, 593, 600–603

F

FEATHERS, 637

Feedback control, 167

Feedforward control, 167

Fidelity, 653

First-order phase transition, 313

Fixed-time strategies, 178

Floating car, 529–532

Floating car data (FCD), 2, 510, 515

Floor field CA, 660

Florida Activity Mobility Simulator (FAMOS), 637

flow interruption interval within wide moving jam, 406

Flow rate, 671

Fluid-dynamic models, 662–663

Force-based models, *see* Acceleration-based models

Forecasting, 617, 618

foreign wide moving jam, 462

Frank–Wolfe algorithm, 88

Free flow, 195, 387

and congested traffic, 202–203

Freely passing buses, 604–606

Free traffic flow, 22

Freeway reconstruction, 528–532

Freeway traffic control

automatic traffic control, 172–173

dynamic speed limits, 180–181

intelligent vehicles and, 187–188

network-oriented, 183–186

network-oriented automatic traffic control, 169–172

ramp metering, 177–180

route guidance, 181–183

sensor technologies, 173–175

traffic flow modeling, 175–177

Friction parameter, 661

F→S transition, 22–23, 31–32, 195, 343, 344, 348, 349, 352, 358

at bottleneck, 395

nucleation nature, 313

F→S→F transitions, 25

F→S→J transitions, 399

Fuel consumption, 501
 acceleration and energy efficiency, 524–528
 empirical measurements, 520–522
 energy matrices, 523–524
 navigation devices, 522
 physical features of traffic phases, 523
 Fundamental diagram, 313, 649, 671
 Future Air traffic management Concept Evaluation Tool
 (FACET) simulation software, 745

G

Gas-kinetic models, 663
 Gas-kinetic traffic flow model, 5
 Gaussian kernel estimation, 675
 General congested pattern (GP), 389, 440, 444
 Generalized forces, 567
 Generalized full velocity difference models, 573
 General Motors (GM) car-following model, 5
 Geographic information system (GIS), 615, 633
 Geographic space, 623
 Gipps-Marksjöes model, 662
 GIS-Interfaced Computational-process modeling for
 Activity Scheduling (GISICAS), 635
 Global positioning system (GPS), 186, 510, 513
 Global telematics protocol (GTP), 531
 GO-pattern/saturation flow pattern, 149
 GP with non-regular pinch region, 470
 Greek fires, August '07, 708

H

Hazard, 701
 types, 704–705
 zone, 701, 703
 Headway, 131, 134, 141
 Heavy traffic congestion, 33, 453
 Herai-Tarui model (HTM), 654
 Hierarchical control, 186
 High occupancy toll (HOT) lanes, 117
 Highway capacity, 346
 of free flow at bottleneck, 196
 meaning of, 10
 Household Activity Pattern Problem (HAPP), 635
 Household Safety Preparedness, and Action Index
 (HSPAI), 707–709, 717
 Household utility models, 639–641
 Human behavior
 goal-oriented, 561
 individuality, 561–563
 intentionality, 563–565
 intermittency, 570–571
 Human intermittent control, 570
 active and passive phases, 570
 main control parameter, 578, 581
 open-loop, 571
 Human perception, 561, 567

Hurricane Katrina, USA, 728
 Hydrodynamic relation, 671

I

Implementation issues, 88–89
 Indifference zone
 in car-following, 344
 in space gap, 359
 in time headway, 349
 Induced traffic breakdown, 22, 30, 62, 196, 204, 211, 216,
 218, 222, 229, 238, 288, 294, 440, 446, 460, 495
 Infinite number, stochastic highway capacities, 349
 IN-Pattern/arrival flow pattern, 149
 Integrated Land Use, Transportation and Environment
 (ILUTE) model, 637
 Intelligent Cruise Control Field Operational Test, 581
 Intelligent driver model (IDM), 232, 573
 Intelligent transportation system (ITS), 2, 7, 26, 80, 131,
 132, 159, 454, 457, 493, 495
 failure of traffic applications, 39–45
 program, 83
 Intelligent Vehicle/Highway System (IVHS), 188
 Intensification of downstream congestion due to upstream
 congestion, 464
 Intentionality, 559
See also Human behavior
 Intermittency, *see* Human behavior
 Intermittency of human actions, 559
 Intermittent control, *see* Human intermittent control
 Inventory, 615

J

Jam/Jamming
 characteristic feature, 388
 and clogging, 677
 suppression effect, 461
 wide moving, 400 (*see also* Wide moving jam)
 Jurisdictions, 623

K

Keep your lane directive, 183
 Kerner-Klenov stochastic microscopic three-phase model,
 48, 313, 315, 370, 501–502
 driver acceleration and deceleration, 260–261
 emergence, history of, 509–510
 over-acceleration effect, 255
 rules of vehicle motion, 252–255
 S→F instability, 257–258
 speed adaptation effect, 256
 speed adaptation vs. over-acceleration, 256
 traffic phases and hypotheses, 506–508
 Z-characteristic for traffic breakdown, 258–260
 Kerner-Klenov-Wolf (KKW) cellular automaton
 (CA) model, 294, 320

Kerner-Klenov-Wolf (KKW) cellular automaton (CA) model (*cont.*)
 driver behavioral assumptions, 295–297
 three-phase traffic flow model, 294–295
 Kerner's three-phase traffic theory, 313, 315, 501, 505, 506, 509, 514, 529, 533, 545
 Kinematic waves (KW) theory, 99
 Knowledge-based methods, 185
 Krauß's stochastic microscopic model, 234

L

Lane changing, 176
 Lane formation, 650, 671, 678
 Language use, 711
 Lattice-gas models, 662
 Lexicographic approach, 626
 Lighthill-Whitham-Richards (LWR) model, 176, 528
 macroscopic traffic flow model, 5
 Limited deceleration model, 335
 Linear piecewise map, 597–598
 Line J, 389, 400
 Link/arc, 79, 90
 Link marginal cost (LMC), 89
 Localized SP (LSP), 440
 Long-term activity and travel planning (LATP), 634
 Long-term prediction with time series, 502

M

Macroscopic criteria for traffic phases [J] and [S], 389, 392, 406
 Macroscopic models, 176, 649, 671–672
 Macroscopic traffic variables, 2
 Main control parameter, 559
 Markovian stochastic process, 298
 Marginal link travel time, 79, 85
 Maximum capacity, mixed flow, 366
 Maximum highway capacity, 22, 36–38
 Maxwell-Boltzmann distribution, 663
 Measurements, traffic flow variables, 15, 502
 Mega-jam, 467
 formation, critical bottleneck strength, 470
 Mesolevel intermittency, 579, 585
 Mesoscopic models, 176
 Metastability, 313
 of free flow, 219
 of free flow with respect to $F \rightarrow S$ transition, 348, 349, 352
 Metropolitan Planning Organizations, 616–617
 Microanalytic Integrated Demographic Accounting System (MIDAS), 634
 Microscopic behavior
 of ACC-vehicle, 379
 of TPACC-vehicle, 379
 Microscopic models, 176, 649, 672
 Microsimulation, 613, 634, 638
 Minimum highway capacity, 23, 36–38

Mitigation, 701
 Mixed traffic flow, 2, 14
 failure of standard approaches for simulations, 350–352
 Model of Action Space in Time Intervals and Clusters (MASTIC), 635
 Model predictive control, 167–168, 185
 Moving SP (MSP), 441
 Moving traffic jam, 9, 23, 196, 388, 391
 Multi-agent systems, 652
 Multilevel regression models, 638–639
 Multi-level traffic control strategies
 hybrid optimization, 160
 real time traffic-adaptive control system (RT-TRACS), 161–163
 signal control and traffic assignment, 161
 system-optimization, 160
 traffic management problem, 159–161
 user-optimization, 160

N

Nagel-Schreckenberg (NaSch) stochastic cellular automaton (CA) model, 233, 315
 Narrow moving jam, 387, 388, 406, 421
 NaSch model, 317
 National Airspace System (NAS), 741, 745
 Natural disaster, 706
 N-curves, 231
 necessary condition for GP existence at heavy bottleneck, 473
 Network and queuing models, 664
 Network capacity, 23, 65–67
 Network-oriented automatic traffic control
 advantages, 184
 blocking, 171
 bottlenecks, 170
 coordination, 184
 prediction, 184
 suboptimal route choice, 170–171
 Network throughput maximization approach, 2, 23, 65–66
 Newell optimal velocity (OV) model, 232
 New research and technology, 615, 623–633
 Non-English speaking households (NESH), 715
 Non-linear map model, 2, 16, 593, 595–597, 609
 Non-linear programming (NLP) problem, 88
 Non-parametric data analysis, 633
 North Queensland, Australia, 722–725
 Nucleation nature
 of $F \rightarrow S$ transition, 313
 of $S \rightarrow J$ transition, 313
 of traffic breakdown, 23, 288, 343, 391
 Nucleus
 moving jam emergence, 388
 for $F \rightarrow S$ transition, 25
 for $S \rightarrow F$ instability, 25
 for $S \rightarrow J$ instability, 421
 for $S \rightarrow J$ transition, 395
 for traffic breakdown, 23–24

O

Objective function, 168
 On-ramp metering
 ALINEA, 44, 267–275
 ANCONA, 44, 45, 264–267
 field trials of, 43–45
 Onset of congestion, 502
 Optimal control, 2, 168, 184
 Optimal flow rate, traffic control method, 40–42
 Optimal reciprocal collision avoidance (ORCA), 657
 Optimal steps model, 662
 Optimal velocity (OV) models, 573, 658
 Optimization policies for adaptive control (OPAC)
 dynamic programming, 154–155
 rolling horizon approach, 156–157
 sequential optimization, 155–156
 VFCOPAC, 153
 virtual fixed cycle strategy, 157–159
 ORIENT, 634
 Origin destination (OD) matrix, 170
 OUT-Pattern/departure flow pattern, 149–150
 Over-acceleration effect, 34, 35, 246
 Over-deceleration effect, 296–297, 420

P

Paradigm shift, traffic and transportation science, 2–3, 6–10
 Path marginal cost (PMC), 89
 Payne's macroscopic traffic flow model, 232
 Peak lanes, 183
 Pedestrian dynamics, 2–3, 649, 672
 acceleration-based models, 654–656
 bottleneck flow, 682–683
 collective effects, 676–680
 controlled experiments, 687–692
 data-driven modelling, 664–665
 density, 675
 density waves, 677
 deterministic vs. stochastic, 652
 discrete vs. continuous, 652
 emergency situation, panic, 679
 flow rate, 673–675
 fundamental diagram, 680–684
 heuristic vs. first-principles, 653
 high vs. low fidelity, 653
 jamming and clogging, 677
 lane formation, 678–679
 macroscopic models, 662–664
 mean speed, 675–676
 microscopic vs. macroscopic, 652
 operational level, 651
 oscillations, 679
 patterns at intersections, 679
 route choice and navigation, 665–666
 rule-based models, 659–662
 rule-based vs. acceleration-based vs. velocity-based, 652–653

simulation frameworks, 666
 on stairs, 684
 stop-and-go waves, 677
 strategic level, 651
 tactical level, 651
 traffic, 4, 16
 uni-and bidirectional, 680
 validation and calibration, 665
 velocity-based models, 656–659
 People's propensity to respond (PPR), 707
 Perception, 711
 Performance-based planning, 617
 Performance function, 168
 Performance index/disutility index, 131, 149, 150
 Personal navigation device (PND), 522
 PESASP, 634
 PETRA, 637
 Phase separation, 330
 Phase transitions, in network, 70–71
 Phenomenology, 559
 See also Human behavior
 Physics of mind, 3, 16, 559
 Piecewise map model, 593, 597–598
 Pinch effect, 296
 Pinch effect, synchronized flow, 389, 418, 445
 Planning applications, 119
 Platoon, 131, 141, 148
 instability, 352
 Poisson distribution, 743
 Policy tools, 619–621
 Portland Daily Activity Schedule Model, 635
 Power law distribution, 744
 Prediction of spatiotemporal congested traffic pattern, 3, 502
 Predictive control, 184
 Predictive feedback controller, 182
 Pre-movement time, 685
 Prevention, 701
 Prism-Constrained Activity-Travel Simulator (PCATS), 635
 Probability
 of traffic breakdown, 346
 of traffic breakdown caused by ACC-vehicles, 375
 of traffic breakdown in mixed flow, 368
 Probe vehicles data, 3, 502
 Probe vehicles, traffic prediction by
 FCD, 515
 freeway sections, 513
 GPS matching, 510, 513
 phase transition, 511, 512
 reconstructed trajectories, 510, 511
 space and time, 510, 511
 synchronized flow patterns reconstruction, 517–519
 upstream fronts of traffic phases, 516–517
 Production systems, 633
 Program Assessment Rating Tool (PART), 616
 Progression models, arterial systems
 arterial signal timings, 141
 closed loops, 141

- Progression models (*cont.*)
 closed network, 141
 mixed-integer linear programming (MILP), 142
 open network, 141
 Propensity and Capacity to Successfully Evacuate (PPCSE), 708
 Prospect theory, 626
- R**
- Radio Data System/Traffic Message Channel (RDS/TMC), 544
 Ramp metering, 177–180
 Ramp-up preparations, 701
 Random time-delayed breakdown, in city traffic
 classical theory, 58–60
 definition, 57
 general features, 60
 metastable under saturated traffic, 60–63
 probabilistic characteristics, 61
 Rational decision making, 625
 RDS/TMC (Radio Data System/Traffic Message Channel), 544
 Real-coded cellular automata (RCA), 662
 Real time traffic-adaptive control system (RT-TRACS)
 degree of adaptability, 161
 levels, 162
 mutual interdependence, 163
 Reciprocal velocity obstacle model (RVO), 657
 Recurring and nonrecurring congestion, 502
 region of wide moving jams of GP, 447
 Representative activity patterns (RAPs), 634
 Resultant safety, 708
 RGB-D sensors, 691
 Risk treatment options, 702
 Road bottleneck, 391
 Road pricing, 79, 117–118
 Roadway capacity, 80
 Ro-Ro passenger ships, 686
 Route guidance, 181–183
 Route/path, 79, 80
 Rule-based models, 649, 659
- S**
- Safe space gap, 242, 363
 Safe speed, 374
 Safe time headway, 356
 Safety-oriented social norm feeds back to individuals [SoEaES1], 703
 Saturation flow, 131, 133
 rate, 24
 Scale-free networks, 741, 744
 SCHEDULER activities, 635
 SCOOT model, 151–153
 Self-averaging local, 562
 Self-driving, 33, 42
 Semi-discrete models, 662
 Semiotics, 706
 Sensor technologies
 advanced, 186–188
 estimation, 175
 intelligent vehicles and traffic control, 187–188
 measurements, 174–175
 Sequence, $F \rightarrow S \rightarrow J$ transitions, 399
 Seven steps to community safety (7SCS), 704
 $S \rightarrow F$ instability, 24, 34–36, 398
 Shetland Islands, 727–728
 Short-term prediction with time series, 502
 Shuttle bus
 dynamics, 16
 systems, 594–596
 Signal control, 132
 and traffic assignment, 161
 Simulation and Assignment in Urban Road Networks (SATURN) approach, 97
 Simulation Model for Activity Resources and Travel (SMART), 635
 Single bus system, 596, 601, 608
 Single intersection, basic model
 average delay per vehicle, 136–138
 cycle time, 138–141
 fixed-time signals, optimum settings, 138
 flow model, 132–133
 fluid traffic model, signalized intersections, 133–134
 green times, 138
 pre-timed signal, continuum model, 134–136
 saturation flow, 133
 $S \rightarrow J$ instability, 296, 387–388, 398, 419, 421, 434, 436
 $S \rightarrow J$ transition, 388, 397, 419, 421
 metastable synchronized flow with respect to, 420, 421
 nucleation nature, 313
 Social-force model (SFM), 572, 655, 656
 Social forces, 572
 Social space, 623, 624
 Social Vulnerability Index (SoVI), 708
 Soft computing, 633
 Space-time clouds, 559, 566
 characterization, 566
 classical approach to self-organization, 567–568
 dynamical trap, 568–570
 illustration, 567
 spatial uncertainty, 567
 spatially separated GPs, 463
 Spatiotemporal congested traffic pattern, 3, 8, 32–33, 502–503
 Kerner's three-phase traffic theory, 533–539
 prediction of, 502
 types of, 508–509
 Speed adaptation effect, 34, 35
 speed adaptation effect
 within 2D states of synchronized flow, 420
 Speed breakdown, 31, 198, 206
See also Traffic breakdown
 Speed disturbances
 by ACC, 363
 by single ACC, 367, 374
 Speed drop, 31
 Speed limits, dynamic, 180–181
 Spontaneous traffic breakdown at bottleneck, 196
 spontaneous wide moving jam emergence, 388
 Standard dynamic traffic assignment, 24, 67–68

- Standard traffic and transportation theories, 3, 5–7, 10, 12, 13
 - Static floor field, 660
 - Static traffic routing and assignment, 79
 - Braess network, 86
 - implementation issues, 88–89
 - system optimum (SO) assignment, 85, 93–96
 - user equilibrium (UE), 85
 - Steady states of traffic flow, 196
 - Stern Report, 714
 - Stochastic highway capacity, 307–308, 349
 - achievements of empirical studies, 210–211
 - classical definition, 209–210
 - failure of classical understanding, 222–223
 - three-phase traffic theory, 38–39
 - Stochastic highway capacity in three-phase theory, 352
 - Stochastic microsimulation, 633
 - Stop-and-go phenomenon, 203
 - Stop-and-go waves, 677
 - String instability of ACC platoon, 364
 - String stability, 48
 - Sustainability implementation research (SIR), 702, 704
 - Sustainable and green visions, 615, 618–622
 - Sustainable transportation, 619
 - Synchronization space gap, 35, 242, 359
 - Synchronization time headway, 356
 - Synchronized flow, 313
 - pattern, 389, 440
 - traffic phase, 24, 27, 31, 196, 388–389, 574
 - synchronized flow, 286, 288, 388
 - Synthetic O-D estimation, 79, 82, 86, 105, 106
 - Synthetic-vision-based steering model, 657, 658
 - System optimum (SO)
 - assignment, 85
 - traffic assignment, 79, 87
- T**
- Temporal scale, 623
 - Terminal Area Forecast (TAF), 742
 - Theoretical fundamentals of transportation science, 24, 63–66
 - Three-phase traffic theory, 3, 22, 24–25, 28, 196–197, 343, 344, 387, 574
 - ACC-vehicle based, 54–56
 - automated driving vehicles, 53–63
 - autonomous driving, 343
 - autonomous driving in framework, 1, 343
 - CA models (*see* Cellular automata (CA) models)
 - congested pattern control approach, 45
 - criticism, 240–242
 - definition, 8
 - description, 315–317
 - driver over-acceleration, 246–248
 - empirical nucleation nature of traffic breakdown, 239–240
 - engineering applications, 15
 - failure of intelligent transportation systems, 39–45
 - highway bottleneck, 34
 - highway capacity, 36–38
 - incommensurability of, 10, 38
 - incommensurability of standard traffic theories, 2
 - main prediction of, 352
 - on-ramp metering, 264–275
 - origin of, 8–9
 - prediction, 343–344
 - prediction of, 10
 - in real world applications, 12
 - S→F instability, 34–36
 - spatiotemporal analysis, 34
 - speed adaptation effect, 244–247
 - stochastic highway capacity, 38–39
 - traffic phases in, 29–31
 - two-dimensional traffic flow states, 242–244
 - Z-characteristic of, 34–36
 - Three traffic phases, 29–31, 387
 - Threshold flow rate, traffic breakdown in mixed flow, 377
 - Tidal flow, 183
 - Time delay, traffic breakdown
 - at highway bottleneck, 25
 - at traffic signal, 25
 - Time-dependent traffic assignment, 79, 97
 - Time series of traffic variables, 503
 - Time-to-interaction (TTI), 657
 - Toronto Area Scheduling model for Household Agents (TASHA), 637
 - TPACC
 - ACC in framework of three-phase theory, 356, 358
 - model, 360
 - strategy, 354
 - Traffic
 - data mining/physics of, 501
 - failure, 7
 - and transportation theories, 5
 - Traffic assignment, 3, 25, 79
 - DTA system (*see* Dynamic traffic assignment (DTA) system)
 - time-dependent, 79, 97
 - Traffic breakdown, 3, 5, 197, 285, 343, 344, 502
 - automated driving vehicles (*see* Automated driving vehicles)
 - classical approaches, 27
 - classical traffic flow models and theories, 33–34
 - congested traffic, 26
 - definition, 286
 - diagram of, 308–310
 - empirical, 22, 27, 31
 - empirical feature of, 5, 6
 - empirical features, 287–288
 - empirical features of, 348
 - empirical nucleation, 31–32
 - empirical nucleation nature, 219–222
 - empirical nucleation nature of, 351
 - empirical proof of nucleation nature, 216–218
 - empirical spatiotemporal traffic phenomena, 32–33
 - empirical spontaneous nucleation, 216
 - free flow conditions, 26
 - at highway bottleneck, 25
 - highway capacity, 36–38
 - local vehicle cluster, 290–292
 - metastability, 7
 - metastable state, 27

- Traffic breakdown (*cont.*)
 - in mixed traffic flow, 48–49
 - model within vehicle cluster, 292–294
 - nucleation and probabilistic features, 285
 - nucleation model, 285–286
 - nucleation nature, 6, 7
 - nucleation rate and mean time delay, 305–306
 - nucleus, 27
 - at on-ramp bottleneck, 30, 51
 - outflow rate from vehicle cluster, 301–303
 - preventing, 63
 - probability, 24, 52, 306–307
 - random sequence of phase transitions, 70–71
 - random time-delayed breakdown, in city traffic, 57–63
 - road bottlenecks, 26
 - S→F instability, 24, 34–36
 - steady states, 304–305
 - stochastic (probabilistic) behavior, 207–208
 - synchronized flow, 27
 - three-phase traffic flow theory, 29–31, 34–45
 - three traffic phases, 203–206
 - time delay and probability, 297–298
 - time delay, at highway bottleneck, 25
 - time delay, at traffic signal, 25
 - Wardrop's UE, 70
 - waves in empirical free flow, 212–215
 - Z-characteristic of, 26, 34–36
- Traffic control technology, 117
- Traffic engineering, 593
- Traffic flow
 - fundamental diagram, 195–196
 - management, 741
 - steady states, 196
 - variables, measurements of, 502
- Traffic flow model, 3, 5, 6, 15, 25, 197, 199
 - classic GM model (*see* Classic General Motors (GM) model)
 - classic LWR model (*see* Classic Lighthill-Whitham-Richards (LWR) model)
 - Kerner-Klenov model (*see* Kerner-Klenov stochastic microscopic three-phase model)
 - Standard traffic flow models, 14
 - three-phase model, 249–252
 - two-phase model, 238–239
- Traffic loading, 79
- Traffic management
 - complex dynamics of, 4
 - complexity of, 4
 - definition, 4
 - vehicular, 5
- Traffic message channel (TMC), 516
- Traffic modeling, 99–102
- Traffic networks
 - breakdown (*see* Traffic breakdown)
 - definition, 80
 - delay estimation, 113
 - driver travel decision behavior modeling, 83–85
 - DTA system (*see* Dynamic traffic assignment (DTA) system)
 - dynamic estimation, measures of effectiveness, 113–116
 - dynamic/time-dependent origin-destination estimation, 105–113
 - dynamic traffic routing, 89–96
 - dynamic travel time estimation, 102–105
 - enhance system performance, 116–118
 - implementation issues, 88–89
 - network capacity measure, 66–67
 - partial stops, 114
 - standard dynamic traffic assignment, 67–68
 - static traffic routing and assignment, 85–86
 - traffic modeling, 99–102
 - transportation areas, 118–119
 - two-route network model, 70, 71
 - user equilibrium vs. system optimum traffic assignment, 87–88
 - vehicle energy consumption and emissions estimation, 114
 - vehicle stops estimation, 113–114
 - Wardrop's user equilibrium, 68
- Traffic pattern
 - expanded pattern, 509
 - general pattern, 508–509
 - prediction approach, 502
 - synchronized pattern, 508
- Traffic phases
 - criteria [S] and [J] for, 389
 - in flow–density plane, 409
 - microscopic criterion, 389
- Traffic prediction, 501
- Traffic prediction, of congested patterns
 - data mining, 504, 533
 - FOTO and ASDA models, 528–529, 547
 - fuel consumption, 519–528
 - in-vehicle traffic prediction, 549–551
 - Kerner's three-phase traffic theory, 506–510
 - long-term traffic prediction, 539–547
 - physics of traffic, 504, 533
 - by probe vehicles, 510–516
 - reconstruction of freeway, 528–532
 - spatiotemporal congested traffic pattern, 508–509
 - spatiotemporal pattern analysis, 533–539
 - synchronized flow patterns reconstruction, at traffic signal, 517–519
 - traffic control centers, 547–549
 - upstream fronts of traffic phase, 516–517
- Traffic-responsive strategies, 178
- Traffic restricting mode, 177
- Traffic routing, 3, 4, 79
 - dynamic (*see* Dynamic traffic routing)
- Traffic science, 593
- Traffic service provider (TSP), 513
- Traffic signal control, 3–5, 14
- Traffic spreading mode, 177
- Traffic stream
 - characteristics of mixed traffic flow, 364–368
 - motion model, 79, 100
- Trajectory, 672
- Trams, complex dynamics of delay, 595, 598–600, 603, 606–609
- TRansportation ANalysis SIMulation System (TRANSIMS), 637

Transportation areas, 118–119
 Transportation modeling and simulation, 613
 Transportation network, 24, 66–67
 Transportation planning applications, 624
 Transportation Research Board (TRB) conference, 616
 Transportation science, theoretical fundamentals, 24, 34–45
 Transportation system
 bus schedules, 594
 combined map model, 603–604
 definition, 593
 delayed map model, 598–600
 dynamic models, 594
 extended circle map model, 600–603
 freely passing buses, 604–606
 increase (decrease) of buses, 608–610
 nonlinear map models, 595–597
 no passing buses, 606–608
 piecewise map model, 597–598
 public transportation system, 594
 shuttle bus systems, 594–596
 single bus, 596, 601, 608
 Transport Protocol Expert Group (TPEG), 516
 TRANSYT, 149
 Travel behavior
 analysis and synthesis, 614
 definition, 614
 demand analysis and prediction, 618, 620–621, 629–631
 dynamic planning practice, 614–618
 evolving modeling paradigm, 633–638
 household utility models, 639–641
 mathematical models, 638
 multilevel regression models, 638–639
 sustainable and green visions, 615, 618–622
 Travel decision behavior, 4
 Travel demand, 4, 613
 and prediction, 618, 620–621, 629–631
 Travel time, 182
 Tversky model, 626
 Two-component approach, 560
 Two-phase traffic flow models, 313–314

U

Uncontrolled manifold, 571
 Upstream jam front, 391
 Urban traffic networks, 7, 14
 analysis of complexity, control system performance, 163–164
 assignment, 131, 132
 controller, 131, 153
 coordination, 131, 141
 cyclic flow profiles, delay-based models, 149–151
 definition, 131
 demand-responsive models, 151–159
 MILP-2 problem, 147–149
 multi-level traffic control strategies, 159–163
 offset, 143

 phases, 131, 132
 progression models, arterial systems, 132, 141–147
 single intersection, basic model, 132–141
 synchronization, 131, 154
 traffic signal control, 132
 User equilibrium (UE) traffic assignment, 85
 vs. system optimum, 87–88
 Wardrop's first principle, 86

V

Variable message signs (VMSs), 80
 Vehicle over-acceleration, 246
 Vehicle-to-infrastructure (V2X) communication, 345
 Vehicular traffic management, 4, 16
 Velocity-based (VB) models, 649, 656–657
 noise, 658–659
 optimal velocity, 658
 velocity obstacle models, 657
 Velocity effect (VE) model, 317–318
 Velocity obstacle models, 657
 Virginia Tech Microscopic Energy and Emission Model (VT-Micro Model), 115–116
 Volition, 560
 Volume (traffic), 131, 142
 Voronoi method, 675
 Vulnerability, 702

W

Wardrop's equilibria, 25, 68
 vs. breakdown minimization (BM) principle, 68–70
 for two-route network, 70–71
 Wardrop's SO states, 25, 68
 Wardrop's user equilibrium, 25, 68
 Wavelet analyses, 545
 Weak congestion, 458
 Webster's model of traffic at signal, 113, 132, 136, 490
 Wide moving jam, 288, 314, 388, 406, 421
 emergence, 503
 flow interruption within, 406
 microscopic criterion, 389
 microscopic criterion for, 389, 402
 microscopic structure, 390
 moving blanks within, 390, 395, 407
 nucleus required, emergence, 388
 in synchronized flow, spontaneous emergence of, 388
 united sequence of, 463
 Wide moving jam traffic phase, 25–26, 197, 388, 574
 Widening SP (WSP), 440
 Wiedemann's psychophysical model, 233
 Woodgate Beach, 725–726

Z

2Z-characteristic, phase transitions, 395
 Z-characteristic, traffic breakdown, 26, 34–36